ENGINEERING-DRIVEN LEARNING APPROACHES FOR BIO-MANUFACTURING AND PERSONALIZED MEDICINE

A Dissertation
Presented to
The Academic Faculty

By

Jialei Chen

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial & Systems Engineering

Georgia Institute of Technology

August 2021

ENGINEERING-DRIVEN LEARNING APPROACHES FOR BIO-MANUFACTURING AND PERSONALIZED MEDICINE

Thesis committee:

Dr. Chuck Zhang, Advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Jianjun (Jan) Shi
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Roshan Joseph, Advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. C. F. Jeff Wu
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Ben Wang
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Arman Sabbaghi
Department of Statistics
Purdue University

Date approved: May 14, 2021

To my parents and my wife, we together made this journey memorable.

ACKNOWLEDGMENTS

I would like to thank my thesis advisors, Professor Chuck Zhang and Professor Roshan Joseph for their precious guidance and continuous support throughout my Ph.D. study. Working with them, as one of their disciples, has been a pleasant and memorable experience.

Professor Zhang has been supportive and has given me the freedom to pursue various projects. I am deeply grateful to him for his insights that directed me to an rewarding research field and helped me sort out research ideas. He has given me lots of helpful advice on how to pursue my academic career path and help me become more mature.

Professor Joseph has taught me innumerable lessons on developing necessary skills and knowledge for a career in academic research. I am grateful to him for holding me to a high research standard and enforcing strict validations for each research result. His kindness, patience, and enthusiasm have also made a significant impact on my personality.

I am also grateful to Professor Ben Wang for his mentorship and the opportunity to collaborate on several projects. I would like to thank Professor Jianjun Shi for his guidance and support, Professor Jeff Wu for his mentorship and suggestions on my future career path, and Professor Arman Sabbaghi for his helpful advice on my thesis. I am also thankful to Dr. Mani Vannan and Professor Vigor Yang for their generous help during my Ph.D. study.

Special thanks to Dr. Simon Mak, Dr. Kan Wang, and Dr. Zhen Qian for the time and energy they have spent with me for many enlightening discussions and enormous help.

I would like to thank all my past and present lab mates: Dr. Jianfeng Shi, Dr. Zih Huei Wang, Dr. Geet Lahoti, Dr. Xuzhou Jiang, Jarod Weber, Arvind Krishna, Hongzhen Tian, Chen Jiang, Michael McCracken, Zhaonan Liu, Daniel Cantrell, Shancong Mou, and Wei Yang for enlightening conversations and research discussions.

My deepest appreciation goes to my parents, Jingbao Chen and Saizhu Han, and to my wife, Yujia Xie. Without your love, encouragement, and most of all patience, I would not be able to achieve this important milestone in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Healthcare problems have tremendous impact on human life. The past two decades have witnessed various biomedical research advances and clinical therapeutic effectiveness, including minimally invasive surgery, regenerative medicine, and immune therapy. However, the development of new treatment methods relies heavily on heuristic approaches and the experience of well-trained healthcare professionals. Therefore, it is often hindered by patient-specific genotypes and phenotypes, operator-dependent post-surgical outcomes, and exorbitant cost. Towards clinically effective and in-expensive treatments, this thesis develops analytics-based methodologies that integrate statistics, machine learning, and advanced manufacturing.

Chapter 1 of my thesis introduces a novel function-on-function surrogate model with application to tissue-mimicking of 3D-printed medical prototypes. Using synthetic metamaterials to mimic biological tissue, 3D-printed medical prototypes are becoming increasingly important in improving surgery success rates. Here, the objective is to model mechanical response curves via functional metamaterial structures, and then conduct a tissue-mimicking optimization to find the best metamaterial structure. The proposed function-on-function surrogate model utilizes a Gaussian process for efficient emulation and optimization. For functional inputs, we propose a spectral-distance correlation function, which captures important spectral differences between two functional inputs. Dependencies for functional outputs are then modeled via a co-kriging framework. We further adopt shrinkage priors to learn and incorporate important physics. Finally, we demonstrate the effectiveness of the proposed emulator in a real-world study on heart surgery.

Chapter 2 proposes an adaptive design method for experimentation under response censoring, often encountered in biomedical experiments. Censoring would result in a significant loss of information, and thereby a poor predictive model over an input domain. For such problems, experimental design is paramount for maximizing predictive power

with a limited budget for expensive experimental runs. We propose an integrated censored mean-squared error (ICMSE) design method, which first estimates the posterior probability of a new observation being censored and then adaptively chooses design points that minimize predictive uncertainty under censoring. Adopting a Gaussian process model with product correlation functions, our ICMSE criterion has an easy-to-evaluate expression for efficient design optimization. We demonstrate the effectiveness of the ICMSE method in an application of medical device testing.

Chapter 3 develops an active image synthesis method for efficient labeling (AISEL) to improve the learning performance in healthcare and medicine tasks. This is because the limited availability of data and the high costs of data collection are the key challenges when applying deep neural networks to healthcare applications. Our AISEL can generate a complementary dataset, with labels actively acquired to incorporate underlying physical knowledge at hand. AISEL framework first leverages a bidirectional generative invertible network (GIN) to extract interpretable features from training images and generate physically meaningful virtual ones. It then efficiently samples virtual images to exploit uncertain regions and explore the entire image space. We demonstrate the effectiveness of AISEL on a heart surgery study, where it lowers the labeling cost by 90% while achieving a 15% improvement in prediction accuracy.

Chapter 4 presents a calibration-free statistical framework for the promising chimeric antigen receptor T cell therapy in fighting cancers. The objective is to effectively recover critical quality attributes under the intrinsic patient-to-patient variability, and therefore lower the cost of cell therapy. Our calibration-free approach models the patient-to-patient variability via a patient-specific calibration parameter. We adopt multiple biosensors to construct a patient-invariance statistic and alleviate the effect of the calibration parameter. Using the patient-invariance statistic, we can then recover the critical quality attribute during cell culture, free from the calibration parameter. In a T cell therapy study, our method effectively recovers viable cell concentration for cell culture monitoring and scale-up.

CHAPTER 1
FUNCTION-ON-FUNCTION KRIGING, WITH APPLICATIONS TO 3D PRINTING OF AORTIC TISSUES

3D-printed medical prototypes, which use synthetic metamaterials to mimic biological tissue, are becoming increasingly important in urgent surgical applications. However, the mimicking of tissue mechanical properties via 3D-printed metamaterial can be difficult and time-consuming, due to the functional nature of both inputs (metamaterial structure) and outputs (mechanical response curve). To deal with this, we propose a novel function-on-function kriging model for efficient emulation and tissue-mimicking optimization. For functional inputs, a key novelty of our model is the spectral-distance (SpeD) correlation function, which captures important spectral differences between two functional inputs. Dependencies for functional outputs are then modeled via a co-kriging framework. We further adopt shrinkage priors on both the input spectra and the output co-kriging covariance matrix, which allows the emulator to learn and incorporate important physics (e.g., dominant input frequencies, output curve properties). Finally, we demonstrate the effectiveness of the proposed SpeD emulator in a real-world study on mimicking human aortic tissue, and show that it can provide quicker and more accurate tissue-mimicking performance compared to existing methods in the medical literature.

1.1 Introduction

Three dimensional (3D) printing is an emerging layer-by-layer additive manufacturing technology, with growing interest in medical applications [1]. This is because 3D-printed prototypes provide precise mimicking of organ shape at an acceptable price and time cost. Such prototypes can be extremely helpful for doctors to practice and be proficient in surgical procedures [2] as well as personalized pre-surgical planning [3]. One limitation is that the

mechanical property (i.e., stress-strain curve) of printed prototypes is completely different from biological tissues [4]. Currently, the state-of-the-art approach is to embed *metamaterial* structure to mimic the desired mechanical property of biological tissue ([5]; see Figure 1.1). However, the optimization for this mimicking may take days or even weeks to perform, due to the *functional* nature of the metamaterial structure. This greatly limits the medical applicability of tissue-mimicking prototypes since surgery timing is a critical factor for outcome success. In this paper, we propose a novel kriging model for emulating functional mechanical response over the design space of functional metamaterial structure, which can be used for efficient tissue-mimicking optimization in practical turnaround times.

There are two key reasons why state-of-the-art tissue-mimicking methods are impractical for *urgent* surgical needs. Firstly, such methods rely solely on both physical experiments and computer experiments, which are expensive and/or time-intensive to run. In particular, a single physical experiment (3D-printing and testing a prototype) takes hours to perform, and a single computer experiment (finite element analysis) requires at least 30 minutes for a reliable mechanical response simulation. Secondly, to optimize for a good structure which mimics the mechanical response of biological tissue, such methods require *many* experimental runs over the design space of functional metamaterial structures. This makes current tissue-mimicking methods prohibitively expensive for urgent surgical applications, where the tissue-mimicking prototype is needed within a day. One strategy (which we adopt) is to train a *surrogate* model (or *emulator*, see [6]) which, given data over the design space, can efficiently *predict* the mechanical response of an untested metamaterial structure. However, due to the expensive nature and functional complexities, it is necessary to integrate the rich physics of the tissue-mimicking problem within the emulator model specification, in order to achieve accurate mimicking in a timely fashion.

The proposed emulator utilizes a technique called *kriging* [7], which models the unknown simulation output via a Gaussian process (GP). Kriging is widely used in computer experiment modeling for its interpolating property, and the fact that both the predictor and

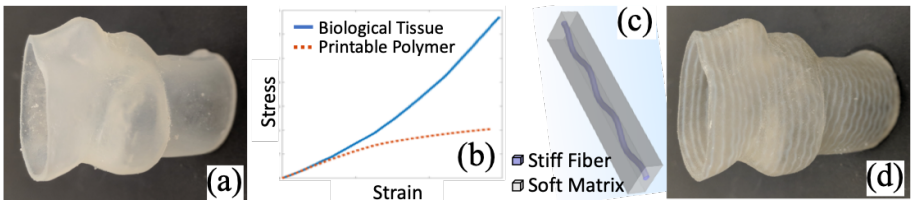

Figure 1.1: *(a) 3D-printed aortic valve (no metamaterial structure), (b) stress-strain curves of biological tissue and printable polymer, (c) a numerical (finite element) simulation case with sinusoidal metamaterial, (d) 3D-printed aortic valve with tissue-mimicking metamaterial.*

its uncertainty have closed-form expressions [6]. The literature on kriging for *functional outputs* typically involves some form of reduced-basis modeling [8, 9, 10, 11] or co-kriging framework [12, 13]. There has also been some work on modeling time series outputs [14]. For *functional inputs*, several techniques have been proposed in functional data analysis literature (see, e.g., [15]), including varying-coefficient models [16] and historical functional linear models [17]. However, the literature on kriging with functional inputs is scarce. For time-series inputs, [18] proposed a kriging model with a covariance function depending on time order. Reduced-basis models were also proposed in [19] and [20]. Such models, however, do not incorporate prior physical knowledge of the tissue-mimicking problem, and can therefore yield poor emulation and mimicking performance given the paucity and functional complexities of the experimental data.

To address this, we introduce in this work a new function-on-function kriging model which integrates an important source of physics: the spectral information of the functional metamaterial structure input. Specifically, we propose a new *spectral-distance* (or SpeD) correlation function, which uses the spectral-distance – the (weighted) Euclidean distance between two functional inputs in spectral domain – to model the process correlation of the GP. This new correlation function captures the appealing property of *translation-invariance*, where two input metamaterial structures which are the same except for a translation shift have the same mechanical properties. We then integrate this within a co-kriging framework for modeling the functional mechanical response output. This emulator-based approach allows for timely and accurate mimicking of biological tissues, and extraction of important

physics (e.g., dominant input frequencies, output curve properties) via sparsity, which broadens the applicability of printed prototypes for urgent surgical procedures.

The paper is structured as follows. Section 1.2 gives an overview of the tissue-mimicking problem. Section 1.3 presents the proposed SpeD emulation model and its shrinkage prior specification. Section 1.4 outlines the algorithm for parameter estimation. Section 1.5 investigates the emulation accuracy, uncertainty quantification, physics extraction and a real-world tissue-mimicking case study. Section 1.6 concludes the work.

1.2 Tissue-mimicking and finite element modeling

We first describe the tissue-mimicking problem (or the metamaterial design problem) and explain the physics of this problem. We then introduce the finite element (FE) analysis as a simulation tool, and provide a brief discussion on experimental design for the FE simulation.

1.2.1 Tissue-mimicking problem

As discussed, 3D-printing technology can print patient-specific prototypes with precise geometry (Figure 1.1 (a)), but the mechanical properties of these printed prototypes can differ greatly from that for true organs (Figure 1.1 (b)). The considered mechanical property is the *stress-strain* curve [21], defined as stress (external tensile load per area) as a function over strain (tensile displacement as a percentage of the specimen length). The stress-strain curve of the biological tissue typically possesses the property of *strain-stiffening*, which means the curve is concave upward (see solid blue line in Figure 1.1 (b)), indicating it becomes stiffer as more load is introduced [4]. However, for 3D-printable material, an opposite property of *strain-softening* is exhibited (see dotted red line in Figure 1.1 (b)) due to the plastic-slipping effect and energy dissipation [22].

To achieve the strain-stiffening property of the biological tissues, one approach is to introduce *metamaterial* structure (i.e., printed enhancement sub-structure) within the prototypes [5]. Figure 1.1 (c) shows an example of a metamaterial with sinusoidal structure.

Here, the *stiffer* enhancement fiber is designed to have a sinusoidal shape, inside the cuboid matrix of a *soft* material. In this work, we treat the structure (or shape) of the enhancement fiber (assumed to have uniform diameter) as the functional input for our SpeD model. Our goal is to mimic the target mechanical property of human tissues, by carefully choosing the shape of the enhancement fiber. Figure 1.1 (d) shows a printed "tissue-mimicking" aortic valve with the optimal metamaterial structure.

1.2.2 Finite element modeling and experimental design

In this work, FE modeling is used to simulate the output stress-strain curve of a given metamaterial structure. FE modeling is frequently used for stress analysis in solid mechanics; it transforms the partial differential equations to their integral form, so that a piece-wise linear formula can be used to approximate the true deformation profile [23]. The key advantage of FE simulations, compared to physical experiments, is that high accuracy can be achieved with no material cost or human error.

Here, FE simulations are performed using COMSOL Multiphysics. The overall size of the metamaterial cuboid (with one enhancement fiber inside) is $20mm$ by $4mm$ by $2mm$, with physics-based quadratic tetrahedral elements for meshing. To compute the stress-strain curve of the metamaterial, one end of the cuboid is fixed while a series of load levels (up to 15% uniaxial deformation) is applied to the other end. The total computation time for one metamaterial is around 30 minutes on 24 Intel Xeon E5-2650 2.20GHz processing cores.

We use a sinusoidal wave structure for designing the training metamaterial structures, as such a form exhibits the best strain-stiffening property from a recent study [5]. The design space has four parameters [24]: the diameter of the enhancement fiber $d \in [0.2, 2]\ mm$, and the amplitude $A \in [0, 1]\ mm$, frequency $\omega \in [0, 0.8]\ mm^{-1}$ and initial phase $\phi \in [0, 2\pi]$ of the sinusoidal wave:

$$I(t) = A\sin(2\pi\omega t + \phi). \tag{1.1}$$

The experimental design adopted for the sinusoidal coefficients is the maximum projection

(MaxPro, [25]) design, which has good space-filling properties on design projections, thereby enabling good predictions from a GP model. Note that the parametric sinusoidal form (1.1) is used only to *generate* data for training the emulator; we will explore a bigger non-parametric input space for *prediction* and tissue-mimicking *optimization*. A total of $n = 58$ metamaterial structures are simulated as the training dataset. An 18-run Sobol' sequence [26] is used as the testing dataset, since it provides a low-discrepancy coverage of the design space, disjoint from the training MaxPro design. Despite the relatively small training dataset ($n = 58$ samples), we show later that the functional stress-strain predictions from the proposed emulator are quite accurate, and provide noticeable improvements over a standard kriging model with four sinusoidal coefficients as inputs.

1.3 Emulation model

We present the proposed emulation model in three parts. First, we introduce the proposed model for functional inputs, using the simplified setting of scalar outputs. We then extend this for functional outputs using a co-kriging structure. Finally, we discuss a prior specification for model parameters which encourages sparsity.

1.3.1 Spectral-distance kriging model

We introduce first the proposed kriging model for functional inputs $I(\cdot) \in \mathcal{I}$, where $\mathcal{I}$ is the functional input space (to be defined later). For simplicity, assume first the case of scalar outputs (functional outputs are introduced next). For the map $y(\cdot) : \mathcal{I} \mapsto \mathbb{R}$ from functional inputs to scalar outputs, we propose the following GP model:

$$y(\cdot) \sim \text{GP}\{\mu, \sigma^2 \rho(\cdot, \cdot)\}, \tag{1.2}$$

where μ is the scalar process mean and σ^2 is the process variance. Here, $\rho(\cdot,\cdot) : \mathcal{I} \times \mathcal{I} \mapsto \mathbb{R}$ is the proposed spectral-distance (SpeD) correlation function, defined as:

$$\rho(I_1(\cdot), I_2(\cdot)) = \text{Corr}\{y(I_1(\cdot)), y(I_2(\cdot))\} = \exp\left(-\text{D}^2\left(\left|\mathcal{F}[I_1(\cdot)]\right|, \left|\mathcal{F}[I_2(\cdot)]\right|; \theta\right)\right). \quad (1.3)$$

Here, $\text{D}(\cdot,\cdot;\theta)$ is a distance function (defined later), $|a\mathbf{i} + \text{b}|$ is the modulus of a complex number $a\mathbf{i} + \text{b}$ (where $\mathbf{i} = \sqrt{-1}$ is the unit imaginary number), and $\mathcal{F}[\cdot] : \mathcal{I} \to \hat{\mathcal{I}}$ is the Fourier transform from the input space of integrable functions, $\mathcal{I} = \{I(\cdot) : \int |I(t)|dt < \infty\}$, to its spectral space $\hat{\mathcal{I}}$. We will use the following definition of a Fourier transform for an input function $I(\cdot)$:

$$\hat{I}(\xi) = \mathcal{F}[I(t)] = \int I(t)e^{-2\pi\mathbf{i}\text{t}\xi}dt, \quad \xi \in \mathbb{R}. \quad (1.4)$$

Similar to the scale-parametrized distance function in the Gaussian correlation (which is widely used for GP emulation of computer experiments, see [6]), we will use the following scale-parametrized l_2 distance function in the *spectral* domain:

$$\text{D}(\left|\mathcal{F}[I_1(\cdot)]\right|, \left|\mathcal{F}[I_2(\cdot)]\right|; \theta) = \left[\int \theta(\xi)\left(\left|\hat{I}_1(\xi)\right| - \left|\hat{I}_2(\xi)\right|\right)^2 d\xi\right]^{1/2}. \quad (1.5)$$

Here, $\theta(\cdot)$ is a weight function in spectral space, with a larger value of $\theta(\xi)$ indicating greater importance of frequency ξ in the SpeD correlation function. In contrast to the standard Gaussian correlation, we assign importance to each *frequency component* of a functional input, rather than to each input *variable*. Plugging (1.5) into (1.3), the SpeD correlation function becomes:

$$\rho(I_1(\cdot), I_2(\cdot)) = \exp\left(-\int \theta(\xi)\left(\left|\hat{I}_1(\xi)\right| - \left|\hat{I}_2(\xi)\right|\right)^2 d\xi\right). \quad (1.6)$$

In our implementation (see Section 1.4), this correlation is computed via a discrete approxi-

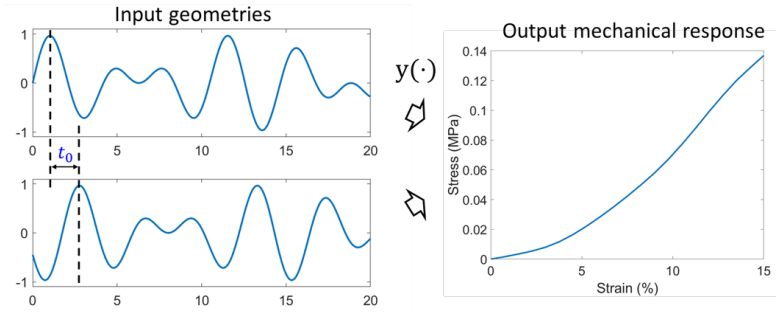

Figure 1.2: *An illustration of the translation-invariance property: for the two input structures which are equivalent up to a translation shift of t_0, their mechanical responses are the same.*

mation of the integral in (1.6).

One advantage of the SpeD correlation function is that it can capture known properties of the tissue-mimicking problem. First, recall that the *translation-shifting* property of Fourier transform [27]: for any $t_0 > 0$, if $I_2(t) = I_1(t - t_0)$, then

$$\hat{I}_2(\xi) = e^{-2\pi \mathbf{i} t_0 \xi} \hat{I}_1(\xi). \tag{1.7}$$

For two metamaterial structures with a shift, i.e., $I_1(t) = I(t)$ and $I_2(t) = I(t - t_0)$, we can then show that their outputs are perfectly correlated, i.e.:

$$\rho(I_1(\cdot), I_2(\cdot)) = \exp\left(- \int \theta(\xi) \left(\left| \hat{I}_1(\xi) \right| - \left| e^{-2\pi \mathbf{i} t_0 \xi} \hat{I}_1(\xi) \right| \right)^2 d\xi \right) = 1. \tag{1.8}$$

We call this the *translation-invariance* property of the SpeD correlation. As illustrated in Figure 1.2, this is a desirable property, since we know from physical knowledge that any translation of the metamaterial structure does not affect the output mechanical response. To contrast, the existing functional input models in Section 1.1 do not enjoy this property. Second, it is known that the stress-strain curve depends largely on frequency ω and amplitude A, but not on initial phase ϕ in the sinusoidal parametrization (1.1) [5, 24]. One can therefore expect that (i) the Fourier frequencies ξ are significant, and (ii) variations in mechanical response are largely due to differences in frequency intensities $|\hat{I}(\xi)|$. The

proposed correlation function (1.6) nicely captures both of these properties.

For our tissue-mimicking problem, the specific choice of the Fourier transform with l_2 distance of the modulus gives an intuitive parametrization of known physical properties. For other applications, the SpeD correlation (1.6) can also be used with other spectral transforms (e.g., wavelet transforms) and other distance metrics (e.g., l_1 distance). The choice of spectral transform and distance should be made on a case-by-case basis, motivated by prior information from the problem at hand.

The following theorem ensures that the SpeD correlation function $\rho(\cdot, \cdot)$ (1.6) is a valid positive semi-definite kernel.

Theorem 1. *The SpeD correlation function $\rho(\cdot, \cdot) : \mathcal{I} \times \mathcal{I} \mapsto \mathbb{R}$ in (1.6), is a positive semi-definite kernel, i.e.:*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \rho(I_i(\cdot), I_j(\cdot)) \geq 0, \tag{1.9}$$

holds for any $n \in \mathbb{N}, c_1, \cdots, c_n \in \mathbb{R}$ and any distinct functions $I_1(\cdot), \cdots, I_n(\cdot) \in \mathcal{I}$.

The proof of Theorem 1 is provided in Appendix A.1. This positive semi-definite property ensures the validity of $\rho(\cdot, \cdot)$ as a proper correlation function to use for GP modeling. Note that $\rho(\cdot, \cdot)$ is not (strictly) *positive-definite*, in that an equality in (1.9) does not imply $c_i = 0$ for all $i = 1, \cdots, n$. This can be seen by setting all input functions $(I_i(\cdot))_{i=1}^n$ to be the same modulo a translation shift; the resulting correlation matrix $[\rho(I_i(\cdot), I_j(\cdot))]_{i=1}^n{}_{j=1}^n$ then becomes a matrix of ones, which is clearly not positive definite. The fact that $\rho(\cdot, \cdot)$ is not positive-definite is not an issue, since for most space-filling designs (including the adopted MaxPro design, see [25]), all training input functions are distinct even after translation shifts.

1.3.2 Spectral-distance co-kriging model

For the tissue-mimicking problem, the output (i.e., the stress-strain curve) is of functional form as well. Below, we generalize the scalar model in Section 1.3.1 to account for functional outputs. Denote the functional input as $I(\cdot) \in \mathcal{I}$ and functional output as $O(\cdot)$, where $O(s)$ is the output stress at strain level s. For our training dataset of $n = 58$ simulated structures, the functional outputs $O_i(\cdot), i = 1, \cdots, n$ are discretized into m levels, yielding output vectors $\mathbf{y}_i \in \mathbb{R}^m$, $i = 1, \cdots, n$. We assume the following SpeD co-kriging model on $\mathbf{y}(\cdot) : \mathcal{I} \mapsto \mathbb{R}^m$:

$$\mathbf{y}(\cdot) \sim \text{GP}\{\boldsymbol{\mu}, \boldsymbol{C}(\cdot, \cdot)\}, \tag{1.10}$$

where $\boldsymbol{\mu} \in \mathbb{R}^m$ is the process mean vector and $\boldsymbol{C}(\cdot, \cdot) : \mathcal{I} \times \mathcal{I} \mapsto \mathbb{R}^{m \times m}$ is the corresponding covariance matrix function.

Consider first the specification of the covariance matrix function $\boldsymbol{C}(\cdot, \cdot)$. Let

$$\boldsymbol{C}(I_1(\cdot), I_2(\cdot)) = \text{Cov}(\mathbf{y}(I_1(\cdot)), \mathbf{y}(I_2(\cdot))) = \rho(I_1(\cdot), I_2(\cdot))\boldsymbol{\Sigma} \quad \text{and} \quad \boldsymbol{\Sigma} \succeq 0. \tag{1.11}$$

Here, $\rho(\cdot, \cdot)$ is the SpeD correlation kernel in (1.6), and $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times m}$ is a symmetric, positive definite co-kriging covariance matrix quantifying correlations between different output levels.

Equation (1.11) implicitly assumes separability in the co-kriging covariance structure. Here, separability means the covariance between output levels observed at different functional inputs can be decomposed as the product of the covariance between output levels and the covariance between functional inputs. This separability assumption is used extensively in the literature for reducing computational complexity [13].

Consider next the specification of mean $\boldsymbol{\mu}$. We assume $\boldsymbol{\mu}$ follows the basis representation:

$$\boldsymbol{\mu} = \mathbf{P}\boldsymbol{\beta}, \tag{1.12}$$

where each column of $\mathbf{P} \in \mathbb{R}^{n\times q}$ represents a pre-specified basis function and $\boldsymbol{\beta} \in \mathbb{R}^q$ denotes its coefficients. This basis representation is similar to the modeling framework of [28, 29]. The choice of basis functions in $\mathbf{P}$ should be guided by prior knowledge on the form of output stress-strain curves. We will describe in Section 1.5.1 a specific parametrization of $\boldsymbol{\mu}$ which incorporates monotonicity information on the stress-strain curve.

Now, we derive the equations for prediction and UQ. Let $\mathbf{y}_{1:n} = \left[\mathbf{y}_1^T, \mathbf{y}_2^T, ..., \mathbf{y}_n^T\right]^T$ denote the vector of functional outputs of the whole training set. Using the conditional distribution formula of the multivariate normal distribution, the discretized functional response $\mathbf{y}_{\text{new}}$ at a new functional input $I_{\text{new}}(\cdot) \in \mathcal{I}$ follows the multivariate normal distribution:

$$\begin{aligned}\mathbf{y}_{\text{new}}|\mathbf{y}_{1:n} \sim \mathcal{N}\Big(&\mathbf{P}_{\text{new}}\boldsymbol{\beta} + (\mathbf{r}_\theta \otimes \boldsymbol{\Sigma})^T \left(\mathbf{R}_\theta^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right)(\mathbf{y}_{1:n} - \mathbf{1}_n \otimes \mathbf{P}\boldsymbol{\beta}),\\ &\boldsymbol{\Sigma} - (\mathbf{r}_\theta \otimes \boldsymbol{\Sigma})^T \left(\mathbf{R}_\theta^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right)(\mathbf{r}_\theta \otimes \boldsymbol{\Sigma})\Big),\end{aligned} \tag{1.13}$$

where $\otimes$ is the Kronecker product, $\mathbf{1}_n$ denotes 1-vector of n elements, $\mathbf{P}_{\text{new}}$ denotes the regression matrix at the new input, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are regression coefficients and co-kriging covariance matrix, $\mathbf{r}_\theta = \left[\rho(I_{\text{new}}(\cdot), I_1(\cdot)), \cdots, \rho(I_{\text{new}}(\cdot), I_n(\cdot))\right]^T$ and $\mathbf{R}_\theta = [\rho(I_i(\cdot), I_j(\cdot))]_{i=1}^n{}_{j=1}^n$. After algebraic manipulations, the posterior mean $\hat{\mathbf{y}}_{\text{new}} = \mathbb{E}\{\mathbf{y}_{\text{new}}|\mathbf{y}_{1:n}\}$ and posterior variance $\text{Var}\{\mathbf{y}_{\text{new}}|\mathbf{y}_{1:n}\}$ can be written in a more concise form:

$$\hat{\mathbf{y}}_{\text{new}} = \mathbb{E}\{\mathbf{y}_{\text{new}}|\mathbf{y}_{1:n}\} = \mathbf{P}_{\text{new}}\boldsymbol{\beta} + \left(\mathbf{r}_\theta^T\mathbf{R}_\theta^{-1} \otimes \mathbf{I}_m\right)(\mathbf{y}_{1:n} - \mathbf{1}_n \otimes \mathbf{P}\boldsymbol{\beta}), \tag{1.14}$$

$$\text{Var}\{\mathbf{y}_{\text{new}}|\mathbf{y}_{1:n}\} = \left(1 - \mathbf{r}_\theta^T\mathbf{R}_\theta^{-1}\mathbf{r}_\theta\right)\boldsymbol{\Sigma}, \tag{1.15}$$

where $\mathbf{I}_m$ denotes an $m \times m$ identity matrix. Equation (1.14) can be used to predict (or emulate) the stress-strain curve for a new metamaterial structure, while Equation (1.15) can be used to construct a confidence band for quantifying the uncertainty of this prediction.

1.3.3 Prior specification

Finally, we provide a prior specification for the model parameters $(\theta(\cdot), \boldsymbol{\Sigma}, \boldsymbol{\beta})$. Consider independent priors on each parameter in $(\theta(\cdot), \boldsymbol{\Sigma}, \boldsymbol{\beta})$. For the weight function $\theta(\cdot)$, we assign independent exponential priors at each frequency ξ, i.e.:

$$\theta(\xi) \overset{i.i.d.}{\sim} \text{Exp}(\lambda_I), \tag{1.16}$$

where λ_I is a rate parameter for the exponential priors. Similar to the Bayesian LASSO [30], the shrinkage prior (1.16) encourages *sparsity* in the maximum a posteriori estimate of $\theta(\cdot)$. This sparsity is desired for two reasons. First, this allows us to identify dominant frequencies in metamaterial structure which influence mechanical response. Second, sparsity in $\theta(\cdot)$ greatly speeds up the tissue-mimicking procedure using the proposed emulator, which is paramount for efficient tissue-mimicking in urgent surgical applications. We note that, in other applications where the time budget allows for a fully Bayesian implementation (see Section 1.4), a spike-and-slab prior [31] could be used.

For the covariance matrix $\boldsymbol{\Sigma}$, we assign the following prior:

$$\pi(\boldsymbol{\Sigma}) \propto \exp(-\lambda_o \|\boldsymbol{\Sigma}^{-1}\|_1). \tag{1.17}$$

Here, λ_o is a rate parameter, and $\|\cdot\|_1$ is the element-wise l_1 norm. The prior (1.17) on $\boldsymbol{\Sigma}$ can be viewed as a shrinkage prior which encourages sparsity on the elements of the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ [32]. This corresponds to the widely-used graphical LASSO [33] method for sparse covariance estimation. For our problem, this sparsity can be used to identify important and interpretable physical couplings in the stress-strain relationship (see Section 1.5.3).

For the regression coefficients $\boldsymbol{\beta}$, we assign a non-informative flat prior $\pi(\boldsymbol{\beta}) \propto 1$, since little information is known on $\boldsymbol{\beta}$ prior to data in our problem. A more informative prior

can be used on $\boldsymbol{\beta}$, if additional domain knowledge is available on the mean trend of the stress-strain curve.

1.4 Parameter estimation

In implementation, the functional inputs $I(\cdot)$ are also discretized to p levels. Let $\boldsymbol{x}_1 = (x_1^k)_{k=0}^{p-1}$ and $\boldsymbol{x}_2 = (x_2^k)_{k=0}^{p-1}$ denote the discretized input vectors for both $I_1(\cdot)$ and $I_2(\cdot)$, respectively. The proposed SpeD kernel $\rho(I_1(\cdot), I_2(\cdot))$ in (1.6) can be approximated as:

$$\rho(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(- \sum_{k=0}^{(p-1)/2} \theta_k \left(\left|\hat{x}_1^k\right| - \left|\hat{x}_2^k\right| \right)^2 \right), \tag{1.18}$$

where $\boldsymbol{\theta} = (\theta_k)_{k=0}^{(p-1)/2}$ is the discretized weight vector, and $\hat{x}^k = \sum_{l=0}^{p-1} x^l e^{-\frac{2\pi \mathbf{i}}{p} lk}$ is the k-th entry of the *discrete* Fourier transform $\hat{\boldsymbol{x}}$ for $\boldsymbol{x}$. Note that $\hat{\boldsymbol{x}}$ is symmetric because $\boldsymbol{x}$ is real-valued [34]; hence, only the first half of $\hat{\boldsymbol{x}}$ is used in (1.18).

With this input discretization, we adopt a maximum a posteriori (MAP) approach for estimating the parameters $(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\Sigma})$. The main reason we prefer MAP over a fully Bayesian approach is computational efficiency, for both parameter estimation and tissue-mimicking optimization. For parameter estimation, a fully Bayesian approach typically requires Markov chain Monte Carlo sampling (MCMC; [35]). Given the complexities of functional inputs and outputs, MCMC sampling can take several days, which is more time-consuming than a single computer experiment run! Furthermore, the primary application of the proposed emulator is for tissue-mimicking optimization, which typically requires *many* evaluations of the emulation predictor. Therefore, it can be *very* time-consuming in a fully Bayesian implementation, since *each* evaluation involves an average over all MCMC samples. In urgent surgical planning, the MAP approach (described next) offers a quicker way to survey the metamaterial design space, which enables timely tissue-mimicking optimization.

From the GP model in (1.10) and (1.13), the MAP estimation of $(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\Sigma})$ boils down to

minimizing the following penalized negative log-posterior [6]:

$$\min_{\boldsymbol{\beta},\boldsymbol{\theta}\geq 0,\boldsymbol{\Sigma}\succeq 0} l_\lambda(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\Sigma}) = \min_{\boldsymbol{\beta},\boldsymbol{\theta}\geq 0,\boldsymbol{\Sigma}\succeq 0} \big[n \log\det \boldsymbol{\Sigma} + m \log\det \mathbf{R}_\theta + \lambda_I \|\boldsymbol{\theta}\|_1 + \lambda_o \|\boldsymbol{\Sigma}^{-1}\|_1 + (\mathbf{y}_{1:n} - \mathbf{1}_n \otimes \mathbf{P}\boldsymbol{\beta})^T \left(\mathbf{R}_\theta^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right) (\mathbf{y}_{1:n} - \mathbf{1}_n \otimes \mathbf{P}\boldsymbol{\beta}) \big]. \tag{1.19}$$

Here, $\mathbf{R}_\theta$ is the correlation matrix in (1.13) with scale parameters $\boldsymbol{\theta}$, and λ_I and λ_o are the rate parameters for the shrinkage priors in Section 1.3.3.

From a regularization perspective, the two prior terms $\lambda_I\|\boldsymbol{\theta}\|_1$ and $\lambda_o\|\boldsymbol{\Sigma}^{-1}\|_1$ in the negative log-posterior (1.19) can equivalently be viewed as penalty terms on $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}^{-1}$, with the rate parameters λ_I and λ_o corresponding to penalization parameters. In this sense, the parameters λ_I and λ_o control the degree of sparsity imposed on $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}^{-1}$, with a larger λ_I (or λ_o) resulting in a sparser estimate of $\boldsymbol{\theta}$ (or $\boldsymbol{\Sigma}^{-1}$), and vice versa. In practice, these penalization parameters can be estimated from the data itself, or specified from the problem at hand. For example, if predictive accuracy of the emulator is the end goal, then λ_I and λ_o can be estimated based on cross-validation techniques [36]. However, if the extraction of important physics is desired, then λ_I and λ_o can be set so that a desired number of physical features can be learned. We will return to this in Section 1.5.3.

Consider now the MAP optimization in (1.19) for fixed $\lambda_I > 0$ and $\lambda_o > 0$. We will use the following blockwise coordinate descent (BCD) optimization algorithm, described below. First, assign initial values for $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$. Next, iterate the following three steps until the convergence is achieved: (i) for fixed GP parameters $\boldsymbol{\theta}$ and regression coefficients $\boldsymbol{\beta}$, compute the correlation matrix $\mathbf{R}_\theta$ and then optimize for covariance matrix $\boldsymbol{\Sigma}$ using the graphical LASSO algorithm [33]; (ii) for fixed $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$, compute $\boldsymbol{\beta}$ using closed-form expressions (see [6] for details); and (iii) for fixed $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, optimize for $\boldsymbol{\theta}$ using the L-BFGS algorithm [37]. The full optimization procedure is provided in Algorithm 1. Since (1.19) is a non-convex optimization problem, the proposed BCD algorithm only converges to a stationary solution [38]. Because of this, we suggest performing multiple runs of Algorithm

Algorithm 1 BCD algorithm for minimizing the penalized negative log-likelihood (19)

- Set initial values $\boldsymbol{\beta} \leftarrow \mathbf{0}_q$, $\boldsymbol{\Sigma} \leftarrow \mathbf{I}_m$ and $\boldsymbol{\theta} \leftarrow \mathbf{1}_p$, and set $\mathbf{Y} \leftarrow [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]^T$

repeat

Optimizing $\boldsymbol{\Sigma}$:

- Set $\mathbf{R}_\theta = \left[\exp\left(-\sum_{k=0}^{(p-1)/2} \theta_q \left(\left|\hat{x}_i^k\right| - \left|\hat{x}_j^k\right|\right)^2\right)\right]_{i=1\,j=1}^{n\quad n}$ with $\hat{x}^k = \sum_{l=0}^{p-1} x^l e^{-\frac{2\pi \mathbf{i}}{p} lk}$
- Set $\boldsymbol{\mu} = \mathbf{P}\boldsymbol{\beta}$
- Set $\mathbf{W}_0 \leftarrow \frac{1}{n}(\mathbf{Y} - \mathbf{1}_n \otimes \boldsymbol{\mu}^T)^T \mathbf{R}_\theta^{-1}(\mathbf{Y} - \mathbf{1}_n \otimes \boldsymbol{\mu}^T) + \lambda_o \cdot \mathbf{I}_m$
- Estimate $\mathbf{W}$ by Graphical LASSO using $\mathbf{W}_0$ as initialization
- Update $\boldsymbol{\Sigma} \leftarrow \mathbf{W}^{-1}$

Optimizing $\boldsymbol{\beta}$:

- Set $\mathbf{S} = (\mathbf{P} \otimes \mathbf{1}_n)^T \left(\mathbf{R}_\theta^{-1} \otimes \mathbf{W}\right)(\mathbf{P} \otimes \mathbf{1}_n)$
- Update $\boldsymbol{\beta} \leftarrow \mathbf{S}^{-1}(\mathbf{P} \otimes \mathbf{1}_n)^T \left(\mathbf{R}_\theta^{-1} \otimes \mathbf{W}\right) \mathbf{y}_{1:n}$

Optimizing $\boldsymbol{\theta}$:

- Update $\boldsymbol{\theta} \leftarrow \operatorname{argmin}_\theta l_\lambda(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\theta})$ with L-BFGS

until $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\theta}$ converge

- **return** $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\theta}$

1 with random initializations for each run, then taking the converged estimates for the run with smallest negative log-likelihood.

1.5 Emulation results

In this section, we present the numerical performance of the proposed model for tissue-mimicking. This is presented in four parts. First, we compare the predictive performance of the proposed SpeD emulation model with two baseline emulation models. Second, we provide a comparison of the uncertainty quantification from these three emulation models. Third, we analyze the physical properties learned via shrinkage priors on $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$. Finally, we demonstrate the usefulness of the fitted model for mimicking human aortic tissue.

1.5.1 Prediction accuracy

As mentioned in Section 1.2.2, the proposed SpeD model is fitted using the training data of $n = 58$ FE simulations. The input function $I(\cdot) \in \mathcal{I}$ is discretized to $p = 81$ parts at $\{0, 0.25, 0.5, \cdots, 20\}$ mm, which we denote as a vector $\boldsymbol{x} \in \mathbb{R}^{81}$. This corresponds to the discretized $\boldsymbol{\theta}$ at frequencies $\{0, 0.05, 0.1, \cdots, 2\}$ mm^{-1}. In the specific tissue-mimicking

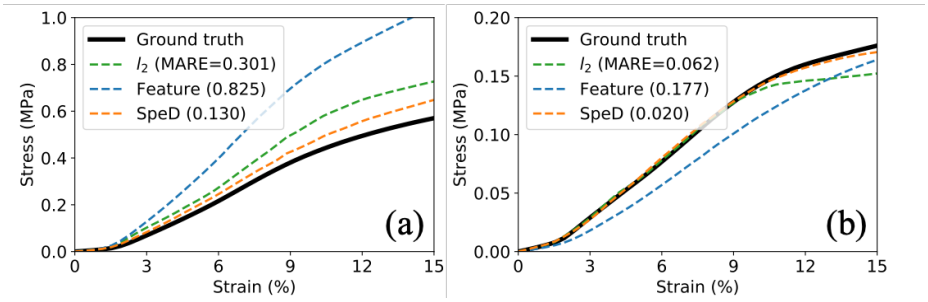

Figure 1.3: *Predicted stress-strain curves for the l_2-distance emulator ("l_2"), feature-based emulator ("Feature"), and the proposed SpeD emulator ("SpeD") on two test metamaterial structures. The corresponding MAREs are included in the legends.*

problem, the diameter of the metamaterial enhancement $d \in \mathbb{R}$ (assumed to be uniform over the whole functional curve $I(\cdot)$) is also important. To account for this extra design variable, we use the following separable correlation function $\rho_s(\cdot,\cdot) : \mathbb{R}^{82} \times \mathbb{R}^{82} \mapsto \mathbb{R}$:

$$\rho_s([d_1, \boldsymbol{x}_1], [d_2, \boldsymbol{x}_2]) = \rho(\boldsymbol{x}_1, \boldsymbol{x}_2) \exp\left(-\theta_d \left(d_1 - d_2\right)^2\right), \tag{1.20}$$

where $\rho(\cdot,\cdot)$ is the discretized SpeD kernel in (1.18) and θ_d is the scale parameter for diameter d. Let $\mathbf{s} \in \mathbb{R}^{m=41}$ denote the vector of strain levels equally spaced from $s = 0$ to $s = 15\%$, and let $\mathbf{y} = O(\mathbf{s}) \in \mathbb{R}^{41}$ be the discretized stress function $O(\cdot)$. Here, the input and output discretization levels are selected heuristically to capture features of the input and output functions: the output functions are quite smooth require less levels, and the input functions are more rugged and require more levels.

From the underlying physics of the stress-strain relationship, it is known that (i) the stress $O(s)$ is always positive, (ii) the stress is zero when the strain is zero (this is known as the free-standing state, see [21]), and (iii) stress-strain curves are typically monotone and non-decreasing, since a larger force is needed to stretch further. To account for (i), a standard log-transformation of stress $O(s)$ is performed prior to modeling and parameter estimation, and the final results are transformed back to ensure the predicted stress is always positive. To account for (ii) and (iii), we choose the basis functions in (1.12) to be $\mathbf{P} = [\mathbf{1}_m, \log(\mathbf{s})]$, along with an additional constraint of $\beta_2 > 0$ to ensure the mean function is monotone

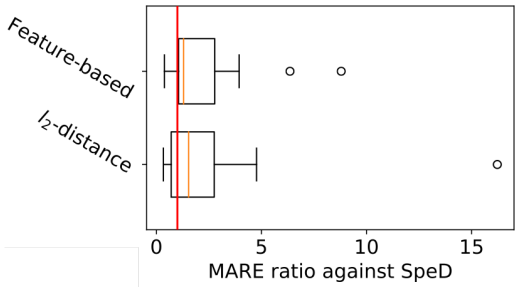

Figure 1.4: *Boxplots of the MARE ratio between the baseline emulators and SpeD emulator on the 18-run test set. The red line marks the MARE ratio of 1.0, where the baseline emulator has the same MARE as the SpeD emulator.*

and non-decreasing. This is equivalent to assuming the mean stress-strain curve takes the following form $O(s) = as^b,\ a, b > 0$, which is a typical parametrization in biomedical literature [1, 24]. This provides a simple and effective way to encourage monotonicity via the mean function specification; one can also extend the shape-constrained GP model in [39] to impose sample path monotonicity, but this is beyond the scope of this work.

For comparison, we also fit two different emulators as baseline methods, using the same dataset. The inputs of the first emulator are the parameters from the sinusoidal wave design $\boldsymbol{x}_p = [d, A, \omega, \phi]^T \in \mathbb{R}^4$, which represents the diameter of the metamaterial fiber, amplitude, period and initial phase of the sinusoidal wave (see Figure 1.1 and Equation (1.1)). This emulator uses a GP model with correlation function:

$$\rho_p(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\sum_{k=1}^{4} \theta_k \left(x_1^k - x_2^k\right)^2\right). \tag{1.21}$$

The same correlation function (with scalar output) is used in [24]. We refer this as the *feature-based* method. The second emulator also assumes a GP model with correlation function:

$$\rho_f(I_1(\cdot), I_2(\cdot)) = \exp\left(-\int \theta(t) \left(I_1(t) - I_2(t)\right)^2 dt\right). \tag{1.22}$$

This correlation (1.22) is essentially the Gaussian correlation function, with distance taken to be the l_2-distance between input functions. A similar correlation function is used in [18]

Table 1.1: *The median MARE of the SpeD emulator and two baseline emulators over the 18-run test set.*

	Median MARE
SpeD	**0.11**
Feature-based	0.19
l_2-distance	0.26

for time-series inputs, with additional dependencies on time order. We refer this as the functional l_2-*distance* method. Both baseline methods assume the same separable co-kriging structure for discretized outputs $\mathbf{y}_{1:n}$, along with MAP parameter estimation.

Predicting stress-strain curve

To test the performance of the proposed emulator, we compare the predictions of stress-strain curves (using Equation (1.14)) for the metamaterial designs from the test set (see Section 1.2.2). Figure 1.3 shows the emulated stress-strain curves for two test metamaterial structures, along with the true stress-strain curve (ground truth) from FE simulations. To quantitatively measure the difference between the predicted and true curves, we use the following mean absolute relative error (MARE) metric:

$$\text{MARE} = \frac{\int_s |O(s) - \hat{O}(s)| ds}{\int_s |O(s)| ds}, \tag{1.23}$$

where s is the strain level, $O(s)$ is the stress at strain s from FE simulation (ground truth), and $\hat{O}(s)$ is the predicted stress from the emulators. The MARE values for the two test cases in Figure 1.3 are included in the legends. Tab 1.1 reports the median MARE values for the three considered emulators, over the whole test set. The proposed SpeD emulator appears to perform very well, in that it achieves noticeably lower median MARE than the two existing emulators. Figure 1.4 shows the boxplots of the MARE ratio between the baseline emulators and the SpeD emulator for the 18 test cases (note that a ratio of 1.0 means the SpeD model yields similar MARE to a baseline model). We see that these ratios are mostly larger than

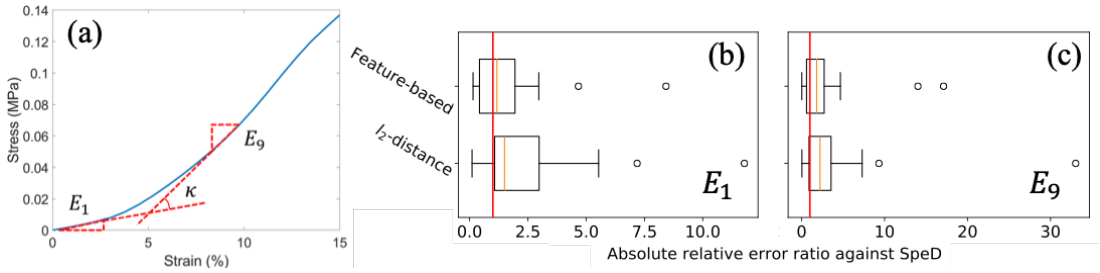

Figure 1.5: *(a) visualizes the three characteristics of mechanical performance: moduli E_1 and E_9, and curvature κ. (b) and (c) show the pairwise absolute relative error for E_1 and E_9 between the two baseline emulators and the SpeD emulator. The red line marks a relative error ratio of 1.0.*

one, which suggests that the proposed emulator is noticeably better in predicting the true stress-strain output curve. This is not surprising, since our model captures known physical properties of the tissue-mimicking problem.

Predicting physical characteristics

In addition to predicting stress-strain curve $O(s)$, engineers are also interested in predicting key physical characteristics. An accurate prediction of these characteristics can be as important as emulating the stress-strain curve itself, because it provides interpretability to the black-box emulation model. Two important physical characteristics of interest are (i) the elastic modulus of the stress-strain curve, and (ii) the classification of material type as strain-stiffening or strain-softening. For (i), the *modulus*, i.e., the slope of the stress-strain curve at different strain levels, can be interpreted as the stiffness or hardness of the material [4]. Here, we are interested in the elastic moduli E_1 and E_9 at strain levels 1% and 9%, respectively, where $E_k = \partial O(s)/\partial s\big|_{s=k\%}$; this allows us to evaluate the elastic moduli prediction over a wide range of strain levels. For (ii), we wish to classify the stress-strain curve as *strain-stiffening* or *strain-softening*; this is particularly important given the goal of mimicking biological tissues (see Section 1.2.1). One way to classify is to use the curvature of the stress-strain curve, which can be approximated by the slope of the two moduli, $\kappa = \partial^2 O/\partial s^2 \approx (E_9 - E_1)/(9\% - 1\%)$. Assuming no fluctuations in $s \in [1, 9]\%$ [21], a positive curvature κ suggests a strain-stiffening property is present (due to increasing

Table 1.2: *The true positive rate, true negative rate, and classification rate of strain-stiffening and strain-softening, for the three considered emulators.*

	SpeD	**Feature-based**	l_2**-distance**
True positive %	12/12=100%	11/12=91.7%	7/12=58.3%
True negative %	6/6=100%	5/6=83.3%	5/6=83.3%
Classification %	18/18=100%	16/18=88.9%	13/18=72.2%

moduli), while a negative κ suggests a strain-softening property is present. Figure 1.5 (a) visualizes these physical characteristics from a stress-strain curve.

We now compare three emulators (SpeD and baselines) for predicting the moduli and material type. The moduli $\hat{E}_1$ and $\hat{E}_9$, computed from the emulated stress-strain curves, are compared with the moduli E_1 and E_9 from FE simulation. Figures 1.5 (b) and (c) show the pairwise absolute relative error $|\hat{E} - E|/E$, between the baselines and the SpeD emulator. We see that most of these ratios are larger than 1.0 in the test set, which shows that the proposed SpeD model outperforms both baseline emulators. For classification, the predicted curvature $\hat{\kappa}$, computed from the emulated curves, are compared with the true curvature κ from FE simulation. Tab 1.2 shows the correct classification rates for the three emulators. The SpeD model has a perfect $18/18 = 100\%$ classification accuracy: it identified the correct strain-softening/-stiffening property for all 18 test structures. On the other hand, the feature-based model and the l_2-distance model achieves only a $16/18 = 88.9\%$ and $13/18 = 72.2\%$ classification accuracy rate, respectively. One reason why the proposed SpeD model can better capture these physical characteristics (compared to existing emulators) is because it directly incorporates the underlying physics via the SpeD correlation function.

1.5.2 Uncertainty quantification

Particularly in healthcare applications, the quantification of predictive uncertainty can be as important as the prediction itself. For the proposed model, Equations (1.14) and (1.15) can be used to construct 90% pointwise highest posterior density predictive intervals (HPD-PIs) for the emulated stress-strain curves. Figure 1.6 shows the 90% HPD-PI for the

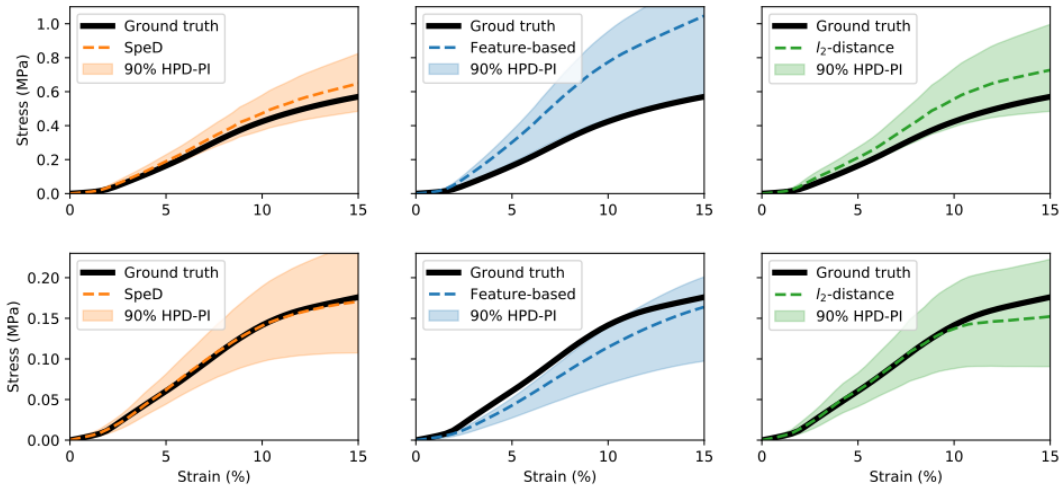

Figure 1.6: *A comparison of the 90% pointwise HPD-PIs for the three emulation models (left: SpeD, middle: feature-based, right: l_2-distance). Different rows are for different test cases.*

three emulation models. Note that there is little predictive uncertainty at low strain, with uncertainty increasing as strain levels increase. This is consistent with the physical intuition in Section 1.5.1: the stress always equals to zero when strain equals zero, i.e., no force at free-standing condition. The increasing uncertainty for higher strain levels may be due to the log-transformation of the functional output.

Comparing the predictive intervals for the three emulators, we see that the proposed SpeD model returns narrower predictive intervals compared to both the l_2-distance model and the feature-based model. This is particular evident for the test case in the top row of Figure 1.6. Moreover, the 90% HPD-PIs of the SpeD emulator covers the true stress-strain curves in $16/18$ of the test cases, whereas the coverage for the feature-based and l_2-distance emulators are only $12/18$ and $14/18$, respectively. For example, the bottom row of Figure 1.6 shows a test case where feature-based emulator fails to cover the true stress-strain curve. Over the whole test set, our SpeD emulator appears to give reliable coverage of the true stress-strain curve, with relatively low predictive uncertainty. The reasons for this may be two-fold: (i) the SpeD correlation captures the physics of the tissue-mimicking problem, which can be viewed as an additional source of data, and (ii) the shrinkage priors on spectral coefficients screens out inert frequencies, which also helps reduce predictive uncertainty. It is worth

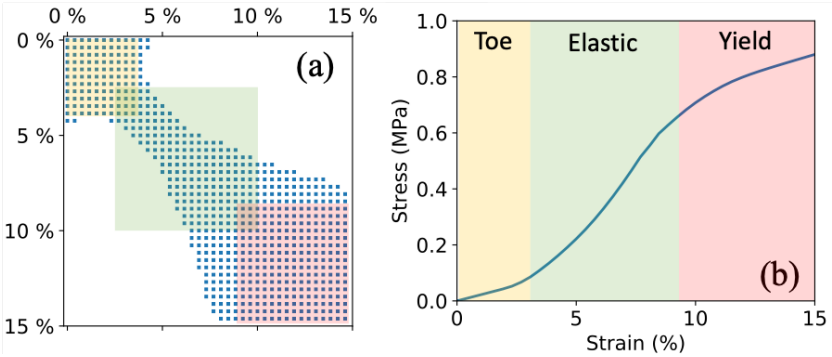

Figure 1.7: *(a) The sparsity pattern visualization of the inverse covariance matrix* $\boldsymbol{\Sigma}^{-1}$ *by graphical LASSO with* 40% *of non-zero entries. The pattern indicates three different regions of the stress-strain curve, colored yellow, green and red. (b) The partition of strain-stress curves for soft materials into toe, elastic, and yield regions up to strain level of* 15%.

noting that the predictive intervals here do not account for parameter uncertainties in the emulator; accounting for such uncertainties would require a fully Bayesian implementation, which would entail much more computational resources.

1.5.3 Learning physics via sparsity

The SpeD emulator also provides a data-driven approach to learn important physics, via the shrinkage priors on both the covariance matrix $\boldsymbol{\Sigma}$ and frequency coefficients $\boldsymbol{\theta}$.

Segmentation of stress-strain curve

We first analyze the important correlations selected by the shrinkage prior on the co-kriging covariance matrix $\boldsymbol{\Sigma}$. Setting the penalty parameter λ_o such that 40% of the entries of $\boldsymbol{\Sigma}^{-1}$ are non-zero, Figure 1.7 (a) visualizes the selected (important) covariances in $\boldsymbol{\Sigma}^{-1}$. Each entry of $\boldsymbol{\Sigma}^{-1}$ represents the corresponding covariance between two stress-strain curve points conditional on all other curve points; note that this covariance quantifies the deviation of the curve from the parametric model $O(s) = as^b$. We see that the stress-strain curve can be roughly segmented into three regions: small strain (from 0% to 3%) with high conditional correlation, medium strain (from 3% to 9%) with moderate conditional correlation and large strain (from 9% to 15%) with high conditional correlation.

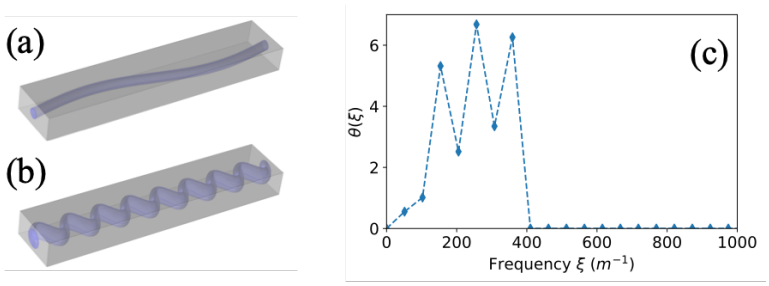

Figure 1.8: *Examples of metamaterial structure with very low (a) or very high (b) frequency. (c) MAP estimates of spectral parameters θ, where medium frequencies are non-zero.*

These three regions suggest a connection to known physical properties in material strength [21, 40], where the mechanical response of the soft bio-mimicking material can also be divided to three regions: the toe region, the elastic region and the yield region (see Figure 1.7 (b)). We see from Figure 1.7 (a) that there are fewer significant conditional correlations in the elastic region compared to the other two regions. One reason for this is that, within the elastic region, the stress-strain curve can be better approximated by the form $O(s) = as^b$ (which corresponds to the choice of basis functions in $\mathbf{P}$). Figure 1.7 (a) also suggests the presence of conditional correlations between the elastic and yield regions. One plausible explanation of this is the migration of strain-stiffening or strain-softening property to straightening.

Learning dominant frequencies

The proposed approach can also learn important frequencies ξ which influence mechanical response, via the shrinkage priors on the weight function $\theta(\xi)$ (see Section 1.3.3). Figure 1.8 (c) shows the MAP estimate of $\theta(\cdot)$ in the spectral space, where the rate parameter λ_I is chosen via cross-validation. We see that $\theta(\cdot)$ shrinks to zero at low and high frequencies, with non-zero estimates only for medium frequencies between $50m^{-1}$ to $400m^{-1}$. For these two endpoint frequencies, Figures 1.8 (a) and (b) show the metamaterial structures with frequencies $\xi \approx 50m^{-1}$ and $\xi \approx 400m^{-1}$, respectively.

The selected frequencies in $\theta(\xi)$ are also in line with the physical understanding of the

problem. For low frequencies (Figure 1.8 (a)), the fluctuation in metamaterial design is too weak to induce any effect on the stress-strain curve, whereas for high frequencies (Figure 1.8 (b)), the resulting strong fluctuation in metamaterial leads to mechanical properties similar to a straight fiber (given nonzero diameter d, see [24]). While it is known that different frequencies affect mechanical response in different ways, a strict law is difficult to find for engineers. Here, our SpeD emulator sheds light on the influential frequencies, i.e., from $50m^{-1}$ to $400m^{-1}$, so those frequencies should be carefully chosen for metamaterial design. We note that these selected frequencies may be sensitive to the choice of experimental design, so further analyses should be taken to confirm such findings from a physics perspective. This identification of important frequencies also allows us to greatly speed up optimization for tissue-mimicking, which we show next.

1.5.4 Mimicking aortic tissue via optimization

We now tackle the motivating task of mimicking the mechanical properties of a target tissue with the proposed emulator. Here, the SpeD model can be used to find a good metamaterial design (both structure $I(\cdot)$ and diameter d) whose stress-strain curve matches the desired mechanical property $\mathbf{y}^*$. This is achieved via the following optimization problem:

$$(d^*, I^*(\cdot)) = \underset{d_{\text{new}}, I_{\text{new}}(\cdot) \in \mathcal{I}}{\operatorname{argmin}} \mathbb{E}\big\{ \|\mathbf{y}([d_{\text{new}}, I_{\text{new}}(\cdot)]) - \mathbf{y}^*\|_2^2 | \mathbf{y}_{1:n} \big\}, \tag{1.24}$$

where $I^*(\cdot)$ is the optimal metamaterial structure, d^* is the optimal fiber diameter, and $\mathbf{y}([d_{\text{new}}, I_{\text{new}}(\cdot)])|\mathbf{y}_{1:n}$ is the conditional (discretized) stress-strain curve in (1.13) with diameter d_{new} and structure $I_{\text{new}}(\cdot)$. In words, equation (1.24) aims to find the optimal metamaterial design whose stress-strain curve from the proposed model (conditional on data) is closest to the target curve $\mathbf{y}^*$ in terms of mean-squared error (MSE).

This MSE criterion can be further decomposed as follows:

$$\|\hat{\mathbf{y}}([d_{\text{new}}, I_{\text{new}}(\cdot)]) - \mathbf{y}^*\|_2^2 + \operatorname{tr}\left(\operatorname{Var}\{\mathbf{y}([d_{\text{new}}, I_{\text{new}}(\cdot)])|\mathbf{y}_{1:n}\}\right). \tag{1.25}$$

Here, $\hat{\mathbf{y}}([d_{\text{new}}, I_{\text{new}}(\cdot)])$ and $\text{Var}\{\mathbf{y}([d_{\text{new}}, I_{\text{new}}(\cdot)])|\mathbf{y}_{1:n}\}$ are the conditional mean and variance of $\mathbf{y}([d_{\text{new}}, I_{\text{new}}(\cdot)])|\mathbf{y}_{1:n}$, respectively, and $\text{tr}(\mathbf{A}) = \sum_i A_{i,i}$ is the trace of the matrix $\mathbf{A}$. The first term can be interpreted as trying to minimize the l_2-norm between the emulated stress-strain curve and the target curve. The second term can be viewed as trying to minimize the predictive variance of the emulated curve. Such a decomposition is quite intuitive, since we wish to find a metamaterial design whose emulated curve matches the desired curve, but also has low predictive uncertainty from the emulation model.

One difficulty in solving (1.24) is that the variable $I(\cdot)$ is *functional* in form, and its discretization $\boldsymbol{x} \in \mathbb{R}^p$, $p = 81$ can be too *high dimensional* to optimize numerically. Here is where the extracted important frequencies from Section 1.5.3 come into play. Let $\hat{\boldsymbol{x}}_{\text{new}} \in \mathbb{R}^7$ denote the seven non-zero Fourier coefficients (see Figure 1.8). Using these coefficients as inputs for optimization (and ignoring the other inert coefficients), we get the following lower-dimensional optimization problem:

$$(d^*, \hat{\boldsymbol{x}}^*) = \underset{d_{\text{new}} \in \mathbb{R}, \hat{\boldsymbol{x}}_{\text{new}} \in \mathbb{R}^7}{\text{argmin}} \mathbb{E}\big\{ \|\mathbf{y}([d_{\text{new}}, \hat{\boldsymbol{x}}_{\text{new}}]) - \mathbf{y}^*\|_2^2 | \mathbf{y}_{1:n} \big\}, \tag{1.26}$$

where $\mathbf{y}([d_{\text{new}}, \hat{\boldsymbol{x}}_{\text{new}}])|\mathbf{y}_{1:n}$ is the conditional random vector taking the frequencies as input. While this problem is non-convex, it is much *lower-dimensional*, and can be effectively optimized using standard quasi-Newton methods (e.g., L-BFGS) and random initializations.

This framework (using the proposed SpeD model) offers significant speeds up for tissue-mimicking over the current state-of-the-art methods. To see why, consider first the optimization of (1.24) using only numerical FE simulations: this requires hundreds of evaluations of the optimization objective function, each of which requires around 30 minutes of computation time. This means tissue-mimicking with only FE simulations can require many days of computation, which is clearly unsuitable for urgent surgical planning [24]. To contrast, each evaluation of the proposed emulator requires only seconds of computation, which greatly speeds up the mimicking process. Furthermore, by exploiting sparsity in

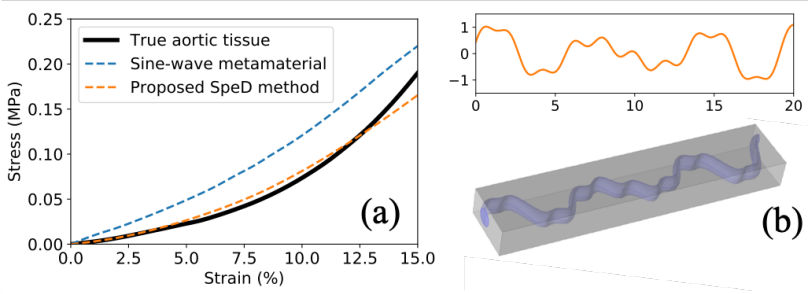

Figure 1.9: *A case study which uses the proposed SpeD model for mimicking human aortic tissue. (a) shows the stress-strain curves of the target aortic tissue (black), the mimicked curve from an existing method (blue) and the curve optimized from the SpeD method (red). (b) shows the optimal metamaterial design from our approach.*

spectral coefficients, the dimension of the optimization problem reduces from 82 to 8 variables. This dimension reduction greatly cuts down on the number of predictions from the emulator, which yields significant reductions in computation time. Such speed-ups are paramount for performing tissue-mimicking in an accurate and timely manner. Section 1.5.5 provides a further comparison of timing.

Figure 1.9 (a) shows the stress-strain curve of a target aortic tissue (in black) from [41], the stress-strain curve from the proposed mimicking procedure (in red), and the curve from an existing mimicking method (in blue) in [5]. The latter method performs mimicking using only the four sinusoidal metamaterial features (see Section 1.2.2). Compared to the existing approach, which has an MARE (see Equation 1.23) of 0.528, the proposed SpeD approach achieves a much smaller MARE of 0.089.

This improved tissue-mimicking performance can be seen in Figure 1.9 (a): the red curve (from the proposed method) closely mimics the desired black curve, whereas the blue curve (from [5]) overestimates stress at all strain levels. In particular, our method gives much better mimicking in small strain regions – this is important in medical applications due to the relatively small strain deformations in the human body.

There are two reasons for this improved performance. First, the existing mimicking approach in [5] is too restrictive, in that it uses only four sinusoidal features and not the full functional form of the input. Second, given a fixed timeframe, the proposed emulation-

Table 1.3: *Computation time for different modeling steps of the proposed emulator, parallelized over 24 processing cores.*

Modeling Step	Computation Time (in minutes)
Parameter estimation	21.72
Prediction & UQ at one input setting	0.01
Tissue-mimicking of a target material	2.69

based approach permits a larger number of objective evaluations via the proposed SpeD model. Figure 1.9 (b) shows the optimal (discretized) metamaterial design $\boldsymbol{x}^*$ from our emulation-based approach, which is clearly not a sinusoidal function. By considering the broader class of functional inputs as well as allowing for more objective evaluations, the proposed method can identify better metamaterial designs for tissue-mimicking.

1.5.5 Computation time

Another appeal of the SpeD emulator is its computational efficiency. Table 1.3 summarizes the computation time required for each step of the emulation process, with timing performed on a parallelized system of 24 Intel Xeon E5-2650 2.20GHz processing cores. The computation time required for parameter estimation (with cross-validation on λ_I) is 21.72 minutes, which is typically performed before the arrival of the patient. Once the model is fit, we can predict for multiple settings very quickly (0.01 minutes for each structure). To contrast, FE simulations require 30 minutes for each structure, and a fully Bayesian implementation of the emulator, which averages over a large amount of MCMC samples (say, 2000), takes at least $0.01 \times 2000 = 20$ minutes per structure. Because of this, our SpeD emulator can effectively perform the tissue-mimicking procedure using only 3 minutes of computation; this greatly improves upon the standard tissue-mimicking approach with only FE simulations, which may require hours or even days to perform [5] with much poorer mimicking performance (see Figure 1.9)! Therefore, the proposed SpeD model can provide effective and personalized pre-surgical practicing and planning [3, 42] with dramatically lower costs, which then mitigates risk in complex surgical procedures.

1.6 Conclusion

We propose in this paper a novel function-on-function Gaussian process emulation model for tackling the challenging tissue-mimicking optimization, under urgent surgical demands. The key challenge is the *functional* input metamaterial structures and the *functional* output mechanical responses.

To address this, the proposed co-kriging model uses a new spectral-distance (SpeD) correlation function, which integrates spectral information by directly modeling the effect of metamaterial frequencies on mechanical response. One appealing feature of this new correlation function is its *translation-invariance* property, which accounts for the fact that two metamaterial structures, which are equivalent modulo a translation shift, have the same mechanical properties. For parameter estimation, we use MAP with shrinkage priors, which identifies key frequencies and thereby reduces the large *functional* input space. This reduction greatly speeds up the tissue-mimicking optimization using the proposed emulator. Applied to a real-world tissue-mimicking study, the proposed SpeD emulator outperforms existing models in (i) emulating and quantifying uncertainty on mechanical response, (ii) extracting meaningful physical insights, and (iii) providing efficient and accurate mimicking performance for human aortic tissue. One direction for future work is the exploration of a more elaborate design method for functional inputs, which may further improve emulation performance. With the development of multi-material 3D-printing technology, this new emulator can play an important role in furthering the impact of 3D-printing in important biomedical applications in surgery planning and healthcare.

CHAPTER 2
ADAPTIVE DESIGN FOR GAUSSIAN PROCESS REGRESSION UNDER CENSORING

A key objective in engineering problems is to predict an unknown experimental surface over an input domain. In complex physical experiments, this may be hampered by response censoring, which results in a significant loss of information. For such problems, experimental design is paramount for maximizing predictive power using a small number of expensive experimental runs. To tackle this, we propose a novel adaptive design method, called the integrated *censored* mean-squared error (ICMSE) method. The ICMSE method first estimates the posterior probability of a new observation being censored, then adaptively chooses design points that minimize predictive uncertainty under censoring. Adopting a Gaussian process regression model with product correlation function, the proposed ICMSE criterion has an easy-to-evaluate expression, which allows for efficient design optimization. We demonstrate the effectiveness of the ICMSE design in two real-world applications on surgical planning and wafer manufacturing.

2.1 Introduction

In many engineering problems, a key objective is to predict an unknown experimental surface over an input domain. However, for complex physical experiments, one can encounter the unfortunate phenomenon of *censoring*, i.e., the experimental response is missing or partially measured. Censoring arises from a variety of practical experimental constraints, including limits in measurement devices, safety considerations of experimenters, and a fixed experimental time budget. Figure 2.1 provides an illustration: experimental censoring typically occurs when the response variable of interest is expensive or dangerous to measure. For predicting the experimental surface, censoring can result in significant loss of information,

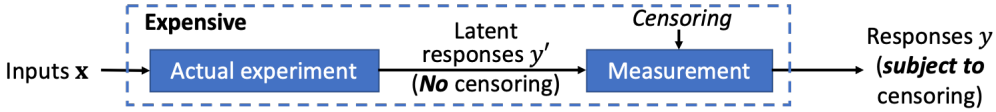

Figure 2.1: *An illustration of response censoring in the measurement process.*

and therefore, poor predictive performance [43]. For example, suppose an engineer wishes to explore how pressure in a nuclear reactor changes under different control settings. Due to safety concerns, experiments are forced to stop if the pressure hits a certain upper limit, leading to censored responses. To further complicate matters, the input region which results in censoring is typically *unknown* prior to experiments, and needs to be estimated from data.

Given the unavoidable and unknown nature of censoring in physical experiments, it is therefore of interest to carefully design experimental runs, to best model the physical system (specifically, the mean function of the experimental surface) via a statistical model. To this end, we present a new integrated *censored* mean-squared error (ICMSE) method, which sequentially selects *physical* experimental runs to minimize predictive uncertainty under *censoring*. ICMSE leverages a Gaussian process model (GP; [44]) – a flexible Bayesian nonparametric model – on the experimental surface, to obtain an easy-to-evaluate design criterion that maximizes GP's predictive power under censoring. We consider two settings of ICMSE, both motivated by real-world applications. The first is a "single-fidelity" ICMSE method for sequentially designing (potentially) censored physical experiments. The second is a "bi-fidelity" ICMSE method for sequentially designing (potentially) censored physical experiments given auxiliary computer simulation data. These two ICMSE methods can easily be extended for a broader range of experimental settings, which we discuss later.

2.1.1 3D-printed aortic valves for surgical planning

The first motivating problem concerns the design of 3D-printed tissue-mimicking aortic valves for heart surgeries. With advances in additive manufacturing [45], 3D-printed medical prototypes [1] play an increasingly important role in pre-surgical studies [3].

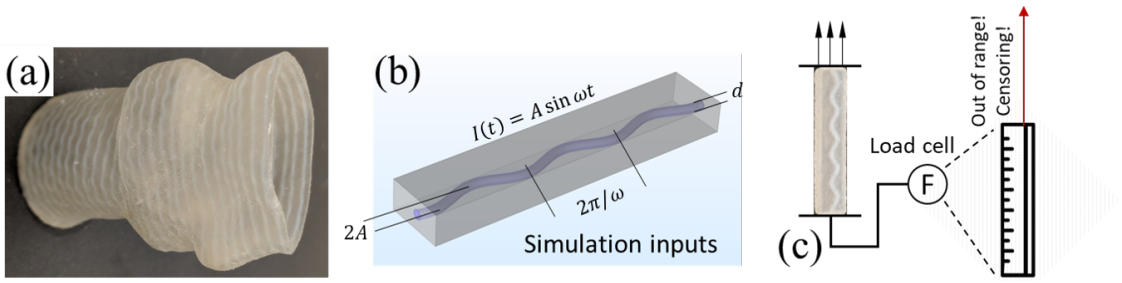

Figure 2.2: *Illustrating the surgical planning application: (a) a 3D-printed aortic valve with enhanced metamaterial, (b) simulation inputs in the computer experiment, (c) visualizing the physical experiment and the measurement censoring of the load cell (labeled "F").*

They are particularly helpful in complicated heart diseases, e.g., aortic stenosis, where 3D-printed aortic valves can be used to select the best surgical option with minimal post-surgical complication [2]. The printed aortic valve (see Figure 2.2(a)) contains a biomimetic substructure: an enhancement polymer (white) is embedded in a substrate polymer (clear); this is known as *metamaterial* [46] in the materials engineering literature. The goal is to understand how the *stiffness* of the metamaterials is affected by the *geometry* of the enhancement polymer (see Figure 2.2(b)).

Using earlier terminology, this is a bi-fidelity modeling problem involving two types of experiments: a pre-conducted database of computer simulations and patient-specific physical experiments. The physical experiments here are very *costly*: we need to 3D print each metamaterial sample, then physically test its stiffness using a load cell. Furthermore, the measurement from physical experiments may be *censored* due to an inherent upper limit of the testing machine. This is shown in Figure 2.2(c): if the metamaterial sample is stiffer than the load cell (i.e., a spring), the experiment is forced to stop to prevent breakage of the load cell. One possible workaround is to use a stiffer load cell, however, it is oftentimes *not* a preferable option: a stiffer load cell with a broader measurement range can be very expensive, costing over a hundred times more than the standard integrated load cells. Here, the proposed ICMSE method can adaptively design experimental runs to maximize the predictive power of a GP model under censoring. The fitted GP model allows for the exploration of different enhancement polymer geometries, which can then be used

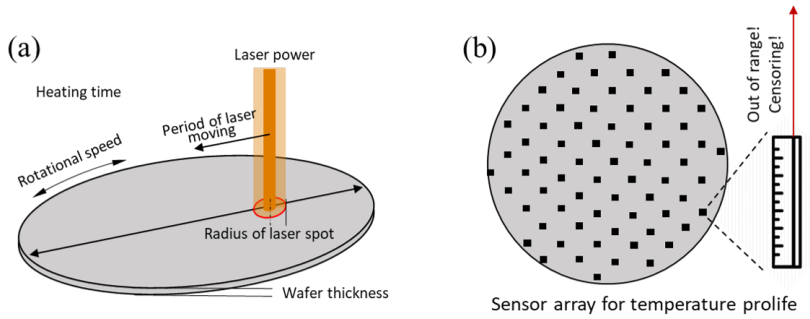

Figure 2.3: *Illustrating the wafer manufacturing application: (a) visualizing the thermal processing procedure with the 6 input parameters, (b) visualizing the measurement censoring of the temperature sensor array.*

for mimicking patient-specific mechanical properties in surgical planning [47]. Our method is particularly valuable in urgent surgical applications, where one can perform only a small number of runs before the actual surgery.

2.1.2 Thermal processing in wafer manufacturing

The second problem considers the design of the semiconductor wafer manufacturing process [48, 49]. Wafer manufacturing involves processing silicon wafers in a series of refinement stages, to be used as circuit chips. Among these stages, thermal processing is one of the most important stages [50], since it facilitates the necessary chemical reactions and allows for surface oxidation. Figure 2.3(a) illustrates the typical thermal processing procedure: a laser beam (in orange) is moved back and forth over a rotating wafer. Here, industrial engineers wish to understand how different process parameters (see Figure 2.3(a)) affect the *minimal* temperature over the whole wafer after heating. The minimal temperature provides information on the completeness of the chemical reactions, and is an important quality measurement in wafer manufacturing [51].

However, laser heating experiments are quite costly, involving high material and operation costs. In industrial settings, the minimal wafer temperature (response variable of interest) is often subject to *censoring*, due to the nature of measurement procedures. This is shown in Figure 2.3(b): the wafer temperature is typically measured by either an array

of temperature sensors or a thermal camera, both of which have upper measurement limits [52]. While more sophisticated sensors exist, they are much more expensive and may lead to tedious do-overs of experiments. The proposed single-fidelity ICMSE method can be used to adaptively design experimental runs that maximize the predictive power of a GP model under censoring. The fitted GP model allows for efficient temperature prediction, which can then be used for quality improvement, real-time control, and other downstream applications.

2.1.3 Literature

GP regression (or *kriging*, see [7]) is widely used as a predictive model for expensive experiments [44], and has been applied in cosmology [53], aerospace engineering [10], healthcare [47], and other applications. The key appeals of GPs are the flexible nonparametric model structure and closed-form expressions for prediction and uncertainty quantification [54]. In the engineering literature, GPs have been used for modeling expensive physical experiments [55], integrating computer and physical experiments [56], and incorporating various constraints [57, 58, 59, 60, 61]. We will adapt in this work a recent censored GP model [62], which integrates censored physical experimental data.

There have been several works in the literature on experimental design under response censoring, see, e.g., [63, 64]. These methods, however, presume a parametric form for the response surface, which may be a dangerous assumption for black-box experiments, hence the recent shift for more nonparametric models such as GPs. Existing design methods for GPs can be divided into two categories – space-filling and model-based designs. Space-filling designs aim to fill empty gaps in the input space; this includes minimax designs [65], maximin designs [66], and maximum projection designs [25]. Model-based designs instead maximize an optimality criterion based on an *assumed* GP model; this includes integrated mean-squared error designs [44] and maximum entropy designs [67]. Such designs can also be implemented sequentially in an adaptive manner, see [68, 69, 70, 71]. Recently, [72] proposed a design method for a heteroscedastic GP model (i.e., under input-dependent

noise); this provides a flexible framework that allows for different correlation functions, closed-form gradients for optimization, and batch sequential implementation.

The above GP design methods, however, do not consider potential response *censoring*. The key challenge in incorporating censoring information is that an experimenter does not know which inputs may lead to censoring prior to experimentation, since the response surface is black-box. The proposed ICMSE method addresses this by leveraging a GP model on the unknown response surface: it first estimates the posterior probability of a potential observation being censored, and then finds design points that minimize predictive uncertainty under censoring. Under product correlation functions, our method admits an easy-to-evaluate design criterion, which allows for efficient sequential sampling. We show that ICMSE can yield considerably improved predictive performance over existing design methods (which do not consider censoring), in both aforementioned motivating applications.

2.1.4 Structure

Section 2.2 presents the ICMSE design method for the single-fidelity setting, with only physical experiment data. Section 2.3 extends the ICMSE method for the bi-fidelity setting, where auxiliary computer simulation data are available. Section 2.4 demonstrates the effectiveness of ICMSE in the two motivating applications. Section 2.5 concludes the work.

2.2 ICMSE design

We now present the ICMSE design method for the single-fidelity setting; a more elaborate bi-fidelity setting is discussed later in Section 2.3. We first review the GP model for censored data, and derive the proposed ICMSE design criterion. We then visualize this via a 1-dimensional (1D) example, and provide some insights.

2.2.1 Modeling framework

We adopt the following model for physical experiments. Let $\mathbf{x}_i \in [0,1]^p$ be a vector of p input variables (each normalized to $[0,1]$), and let y_i' be its latent response from the physical experiment *prior to* potential censoring (see Figure 2.1). We assume:

$$y_i' = \xi(\mathbf{x}_i) + \epsilon_i, \quad i = 1, 2, \cdots, n, \tag{2.1}$$

where $\xi(\mathbf{x}_i)$ is the mean of the latent response y_i' at input $\mathbf{x}_i$, and ϵ_i is the corresponding measurement error. Since $\xi(\cdot)$ is unknown, we further assign to it a GP prior with mean μ_ξ, variance σ_ξ^2, and correlation function $R_{\boldsymbol{\theta}_\xi}(\cdot,\cdot)$ with parameters $\boldsymbol{\theta}_\xi$. This is denoted as:

$$\xi(\cdot) \sim \text{GP}\left(\mu_\xi, \sigma_\xi^2 R_{\boldsymbol{\theta}_\xi}(\cdot,\cdot)\right). \tag{2.2}$$

The experimental noise $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ is assumed to be i.i.d. normally distributed, and independent of $\xi(\cdot)$.

For simplicity, we consider only the case of right-censoring below, i.e., censoring of the response only when it exceeds some *known* upper limit (this is the setting for both motivating applications). All equations and insights derived in the paper hold analogously for the general case of interval censoring, albeit with more cumbersome notation. Suppose, from n experiments, n_o responses are observed without censoring, and n_c responses are right-censored at limit c, where $n_o + n_c = n$. The training set experimental data can then be written as the set $\mathcal{Y}_n = \{\mathbf{y}_o, \mathbf{y}_c' \geq \mathbf{c}\}$, where $\mathbf{y}_o$ is a vector of *observed* responses at inputs $\mathbf{x}_o = \mathbf{x}_{1:n_o} = \{\mathbf{x}_1, \cdots, \mathbf{x}_{n_o}\}$, $\mathbf{y}_c'$ is the latent response vector for inputs in censored regions $\mathbf{x}_c = \mathbf{x}_{(n_o+1):n}$ prior to censoring, and $\mathbf{c} = [c, \cdots, c]^T$ is the vector of the right-censoring limit. Assuming known model parameters, a straightforward adaptation of the equations (11) and (12) in [62] gives the following expressions for the conditional mean and variance

of $\xi(\mathbf{x}_{\text{new}})$ at new input $\mathbf{x}_{\text{new}}$:

$$\hat{\xi}(\mathbf{x}_{\text{new}}) = \mathbb{E}[\xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n] = \mu_\xi + \boldsymbol{\gamma}_{n,\text{new}}^T \boldsymbol{\Gamma}_n^{-1} \left([\mathbf{y}_o, \hat{\mathbf{y}}_c]^T - \mu_\xi \cdot \mathbf{1}_n\right), \tag{2.3}$$

$$s^2(\mathbf{x}_{\text{new}}) = \text{Var}[\xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n] = \sigma_\xi^2 - \boldsymbol{\gamma}_{n,\text{new}}^T (\boldsymbol{\Gamma}_n^{-1} - \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\Sigma} \boldsymbol{\Gamma}_n^{-1}) \boldsymbol{\gamma}_{n,\text{new}}. \tag{2.4}$$

Here, $\boldsymbol{\Gamma}_n = \sigma_\xi^2 [R_{\boldsymbol{\theta}_\xi}(\mathbf{x}_i, \mathbf{x}_j)]_{i=1}^n{}_{j=1}^n + \sigma_\epsilon^2 \mathbf{I}_n$, $\boldsymbol{\gamma}_{n,\text{new}} = \sigma_\xi^2 \big[R_{\boldsymbol{\theta}_\xi}(\mathbf{x}_1, \mathbf{x}_{\text{new}}), \cdots, R_{\boldsymbol{\theta}_\xi}(\mathbf{x}_n, \mathbf{x}_{\text{new}})\big]^T$, $\mathbf{1}_n$ is a one-vector of length n, and $\mathbf{I}_n$ is an $n \times n$ identity matrix. Furthermore, $\hat{\mathbf{y}}_c = \mathbb{E}[\mathbf{y}_c'|\mathcal{Y}_n]$ is the expected response for the latent vector $\mathbf{y}_c'$ given the dataset $\mathcal{Y}_n$, $\boldsymbol{\Sigma}_c = \text{Var}[\mathbf{y}_c'|\mathcal{Y}_n]$ is its conditional variance, and $\boldsymbol{\Sigma} = \text{diag}(\mathbf{0}_{n_o}, \boldsymbol{\Sigma}_c)$. The computation of these quantities will be discussed later in Section 2.3.3. The conditional mean (2.3) is used to predict the mean experimental response at an untested input $\mathbf{x}_{\text{new}}$, and the conditional variance (2.4) is used to quantify predictive uncertainty.

In the case of no censoring (i.e., $\mathcal{Y}_n = \{\mathbf{y}_o\}$), equations (2.3) and (2.4) reduce to:

$$\hat{\xi}(\mathbf{x}_{\text{new}}) = \mathbb{E}[\xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n] = \mu_\xi + \boldsymbol{\gamma}_{n,\text{new}}^T \boldsymbol{\Gamma}_n^{-1} (\mathbf{y}_o - \mu_\xi \cdot \mathbf{1}_n), \quad \text{and} \tag{2.5}$$

$$s^2(\mathbf{x}_{\text{new}}) = \text{Var}[\xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n] = \sigma_\xi^2 - \boldsymbol{\gamma}_{n,\text{new}}^T \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_{n,\text{new}}. \tag{2.6}$$

These are precisely the conditional mean and variance expressions for the standard GP regression model [54], which is as expected.

2.2.2 Design criterion

Now, given data $\mathcal{Y}_n$ from n experiments (n_o of which are observed exactly, n_c of which are censored), we propose a new design method that accounts for the posterior probability of a potential observation being censored. Let $\mathbf{x}_{n+1}$ be a potential next input for experimentation, Y_{n+1}' be its latent response *prior to* censoring, and $Y_{n+1} = Y_{n+1}'(1 - \mathbb{1}_{\{Y_{n+1}' \geq c\}}) + c\mathbb{1}_{\{Y_{n+1}' \geq c\}}$ be its corresponding observation *after* censoring, with $\mathbb{1}_{\{\cdot\}}$ denoting the indicator function.

The proposed method chooses the next input $\mathbf{x}_{n+1}^*$ as:

$$\begin{aligned}\mathbf{x}_{n+1}^* &= \underset{\mathbf{x}_{n+1}}{\operatorname{argmin}}\, \text{ICMSE}(\mathbf{x}_{n+1}) \\ &:= \underset{\mathbf{x}_{n+1}}{\operatorname{argmin}} \int_{[0,1]^p} \mathbb{E}_{Y_{n+1}|\mathcal{Y}_n}\left[\text{Var}(\xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n, Y_{n+1})\right]\, d\mathbf{x}_{\text{new}}.\end{aligned} \tag{2.7}$$

The design criterion $\text{ICMSE}(\mathbf{x}_{n+1})$ can be understood in two parts. First, the term $\text{Var}(\xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n, Y_{n+1})$ quantifies the predictive variance (i.e., mean-squared error, MSE) of the mean response at an untested input $\mathbf{x}_{\text{new}}$, given both the training data $\mathcal{Y}_n$ and the potential observation Y_{n+1}. This is a reasonable quantity to minimize for design, since we wish to find which new input $\mathbf{x}_{n+1}$ can minimize predictive uncertainty. Second, note that this MSE term cannot be used directly as a criterion, since it depends on the potential observation Y_{n+1}, which is yet to be observed. One way around this is to take the conditional expectation $\mathbb{E}_{Y_{n+1}|\mathcal{Y}_n}[\cdot]$ (more on this below). Finally, the integral over $[0,1]^p$ yields the average predictive uncertainty over the entire design space.

The proposed criterion in (2.7) can be viewed as an extension of the sequential integrated mean-squared error (IMSE) design [68, 54] for the censored response setting. Assuming no censoring (i.e., $\mathcal{Y}_n = \{\mathbf{y}_o\}$), the sequential IMSE design chooses the next input $\mathbf{x}_{n+1}^*$ by minimizing:

$$\min_{\mathbf{x}_{n+1}} \text{IMSE}(\mathbf{x}_{n+1}) := \min_{\mathbf{x}_{n+1}} \int_{[0,1]^p} \text{Var}(\xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n, Y'_{n+1})\, d\mathbf{x}_{\text{new}}. \tag{2.8}$$

Note that, in the *uncensored* setting, the MSE term $\text{Var}(\xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n, Y'_{n+1})$ in (2.8) does *not* depend on the potential observation Y'_{n+1}, which allows the criterion to be easily computed in practice. However, in the *censored* setting at hand, not only does this MSE term *depend* on Y'_{n+1}, but such an observation may not be directly observed due to censoring. The conditional expectation $\mathbb{E}_{Y_{n+1}|\mathcal{Y}_n}[\cdot]$ in (2.7) addresses this by accounting for the posterior probability of censoring in Y'_{n+1}.

One attractive feature of the ICMSE criterion (2.7) is that it will be *adaptive* to the

experimental responses from data. The criterion (2.7) inherently hinges on whether the potential observation Y_{n+1} is censored (i.e., $Y'_{n+1} \geq c$) or not (i.e., $Y'_{n+1} < c$), but this censoring behavior needs to be estimated from experimental data. Viewed this way, the ICMSE criterion can be broken down into two steps: it (i) estimates the posterior probability of a new observation being censored from data, and then (ii) samples the next point that minimizes the *average* predictive uncertainty under censoring. We will show how our method adaptively incorporates the posterior probability of censoring Y_{n+1} for sequential design, in contrast to the existing IMSE method (2.8).

No censoring in training data

To provide some intuition, consider a simplified scenario with no censoring in the *training* set, i.e., $\mathcal{Y}_n = \{\mathbf{y}_o\}$ (censoring may still occur for the new Y_{n+1}). In this case, the following theorem gives an explicit expression for the ICMSE criterion.

Theorem 2. *Suppose there is no censoring in training data, i.e.,* $\mathcal{Y}_n = \{\mathbf{y}_o\}$. *Then the ICMSE criterion* (2.7) *has the explicit expression:*

$$\text{ICMSE}(\mathbf{x}_{n+1}) = \int_{[0,1]^p} \sigma^2_{\text{new}} - h_c(\mathbf{x}_{n+1})\rho^2_{\text{new}}(\mathbf{x}_{n+1})\sigma^2_{\text{new}}\, d\mathbf{x}_{\text{new}}, \tag{2.9}$$

$$where \quad h_c(\mathbf{x}_{n+1}) = h(z_c) = \Phi(z_c) - z_c\phi(z_c) + \frac{\phi^2(z_c)}{1-\Phi(z_c)}, \quad z_c = \frac{c-\mu_{n+1}}{\sigma_{n+1}}.$$

Here, $\sigma^2_{\text{new}} = \text{Var}[\xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n]$, $\rho_{\text{new}}(\mathbf{x}_{n+1}) = \text{Corr}[\xi(\mathbf{x}_{n+1}), \xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n]$, $\mu_{n+1} = \mathbb{E}[\xi(\mathbf{x}_{n+1})|\mathcal{Y}_n]$, *and* $\sigma^2_{n+1} = \text{Var}[\xi(\mathbf{x}_{n+1})|\mathcal{Y}_n]$ *follow from* (2.5) *and* (2.6). $\phi(\cdot)$ *and* $\Phi(\cdot)$ *are the probability density and cumulative distribution functions for the standard normal distribution.*

In words, μ_{n+1} is the predictive mean at $\mathbf{x}_{n+1}$ given data $\mathcal{Y}_n$, σ^2_{n+1} and σ^2_{new} are the predictive variances at $\mathbf{x}_{n+1}$ and $\mathbf{x}_{\text{new}}$, respectively, and $\rho_{\text{new}}(\mathbf{x}_{n+1})$ is the posterior correlation between $\xi(\mathbf{x}_{n+1})$ and $\xi(\mathbf{x}_{\text{new}})$. Note that the p-dimensional integral in (2.9) can also be efficiently

computed in practice; we provide more discussion later in Corollary 1. The proof of this theorem can be found in Appendix B.1.2.

To glean intuition from the criterion (2.9), we compare it with the existing sequential IMSE criterion (2.8). Under no censoring in training data (i.e., $\mathcal{Y}_n = \{\mathbf{y}_o\}$), (2.8) can be rewritten as:

$$\text{IMSE}(\mathbf{x}_{n+1}) = \int_{[0,1]^p} \sigma^2_{\text{new}} - \rho^2_{\text{new}}(\mathbf{x}_{n+1})\sigma^2_{\text{new}}\, d\mathbf{x}_{\text{new}}. \tag{2.10}$$

Comparing (2.10) with (2.9), we note a key distinction in the ICMSE criterion: the presence of $h_c(\mathbf{x}_{n+1}) = h(z_c)$, where z_c is the normalized right-censoring limit under the posterior distribution at $\mathbf{x}_{n+1}$. We call $h(\cdot)$ the *censoring adjustment* function. Figure 2.4 visualizes $h(z_c)$ for different choices of z_c. Consider first the case of z_c large. From the figure, we see that $h(z_c) \to 1$ as $z_c \to \infty$, in which case the proposed ICMSE criterion (2.9) reduces to the standard IMSE criterion (2.10). This makes sense intuitively: a large value of z_c (i.e., a high right-censoring limit) means that a new observation at $\mathbf{x}_{n+1}$ has little posterior probability of being censored at c. In this case, the ICMSE criterion (which minimizes predictive variance *under* censoring) should then reduce to the IMSE criterion (which minimizes predictive variance *ignoring* censoring). Consider next the case of z_c small. From the figure, we see that $h(z_c) \to 0$ as $z_c \to -\infty$, and the proposed criterion (2.9) reduces to the integral of σ^2_{new}. Again, this makes intuitive sense: a small value of z_c (i.e., a low right-censoring limit) means a new observation at $\mathbf{x}_{n+1}$ has a high posterior probability of being censored. In this case, the ICMSE criterion reduces to the predictive variance of the testing point $\mathbf{x}_{\text{new}}$ given only the first n training data points, meaning a new design point at $\mathbf{x}_{n+1}$ offers little reduction in predictive variance. Viewed this way, the proposed ICMSE criterion modifies the standard IMSE criterion by accounting for the posterior probability of censoring via the censoring adjustment function $h(z_c)$.

Equation (2.9) also reveals an important trade-off for the proposed design under cen-

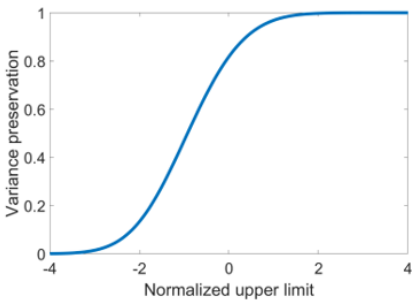

Figure 2.4: *Visualizing the censoring adjustment function $h(z_c)$, where z_c is the normalized right-censoring limit.*

soring. Consider first the standard IMSE criterion (2.10), which minimizes predictive uncertainty under no censoring. Since the first term σ^2_{new} does not depend on the new design point $\mathbf{x}_{n+1}$, this uncertainty minimization is achieved by maximizing the second term $\rho^2_{\text{new}}(\mathbf{x}_{n+1})\sigma^2_{\text{new}}$. This can be interpreted as the variance reduction from observing Y'_{n+1} [73]. Consider next the proposed ICMSE criterion (2.9), which maximizes the term $h(z_c)\rho^2_{\text{new}}(\mathbf{x}_{n+1})\sigma^2_{\text{new}}$. This can further be broken down into (i) the maximization of variance reduction term $\rho^2_{\text{new}}(\mathbf{x}_{n+1})\sigma^2_{\text{new}}$, and (ii) the maximization of the censoring adjustment function $h(z_c)$. Objective (i) is the same as for the standard IMSE criterion – it minimizes predictive uncertainty assuming no response censoring. Objective (ii), by maximizing the censoring adjustment function $h(z_c)$, aims to minimize the posterior probability of the new design point being censored. Putting both parts together, the ICMSE criterion (2.9) features an important trade-off: it aims to find a new design point that jointly minimizes predictive uncertainty (in the absence of censoring) and the posterior probability of being censored.

Censoring in training data

We now consider the general case of censored training data $\mathcal{Y}_n = \{\mathbf{y}_o, \mathbf{y}'_c \geq \mathbf{c}\}$. The following theorem gives an explicit expression for the ICMSE criterion.

Theorem 3. *Given the censored data* $\mathcal{Y}_n = \{\mathbf{y}_o, \mathbf{y}_c' \geq \mathbf{c}\}$*, we have:*

$$\text{ICMSE}(\mathbf{x}_{n+1}) = \int_{[0,1]^p} \sigma^2_{\text{new}} - \boldsymbol{\gamma}^T_{n+1,\text{new}} \boldsymbol{\Gamma}^{-1}_{n+1} \mathbf{H}_c(\mathbf{x}_{n+1}) \boldsymbol{\Gamma}^{-1}_{n+1} \boldsymbol{\gamma}_{n+1,\text{new}} \, d\mathbf{x}_{\text{new}}, \qquad (2.11)$$

where $\sigma^2_{\text{new}} = \text{Var}[\xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n]$*, and* $\boldsymbol{\gamma}_{n+1,\text{new}}$ *and* $\boldsymbol{\Gamma}_{n+1}$ *follow from* (2.3) *and* (2.4)*. The matrix* $\mathbf{H}_c(\mathbf{x}_{n+1})$ *has an easy-to-evaluate expression given in Appendix B.1.3.*

Here, σ^2_{new} is the predictive variance at point $\mathbf{x}_{\text{new}}$ conditional on the data $\mathcal{Y}_n$. The full expression for $(n+1) \times (n+1)$ matrix $\mathbf{H}_c(\mathbf{x}_{n+1})$, while easy-to-evaluate, is quite long and cumbersome; this expression is provided in Appendix B.1.3. The key computation in calculating $\mathbf{H}_c(\mathbf{x}_{n+1})$ is evaluating several orthant probabilities from a multivariate normal distribution. The proof for this theorem can be found in Appendix B.1.3. Section 2.3.3 and Appendix B.3 provide further details on computation.

While this general ICMSE criterion (2.11) is more complex, its interpretation is quite similar to the earlier criterion – its integrand contains a posterior variance term conditional on data $\mathcal{Y}_n$, and a variance reduction term from the potential observation Y_{n+1}. The matrix $\mathbf{H}_c(\mathbf{x}_{n+1})$ on the variance reduction term serves a similar purpose to the censoring adjustment function. A large value of $\mathbf{H}_c(\mathbf{x}_{n+1})$ (in a matrix sense) suggests a low posterior probability of censoring for a new point $\mathbf{x}_{n+1}$, whereas a small value suggests a high posterior probability of censoring. This again results in the important trade-off for sequential design under censoring: the proposed ICMSE criterion aims to find the next design point which not only (i) minimizes predictive uncertainty of the fitted model in the absence of censoring, but also (ii) minimizes the posterior probability that the resulting observation is censored. The posterior probability is adaptively learned from the training data, and is not considered by the standard IMSE criterion.

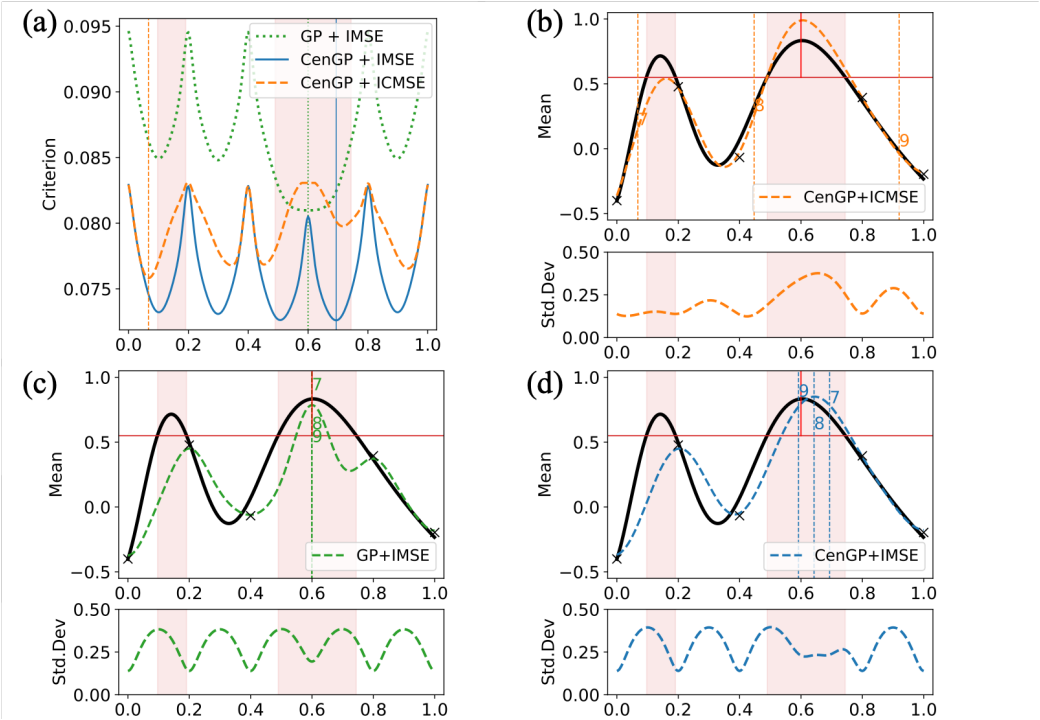

Figure 2.5: *A 1D illustrative example* (2.12)*: (a) shows the design criteria of the next run x_7 for the three considered methods. (b), (c), and (d) show the 3 sequential runs (x_7^*, x_8^*, x_9^*) using CenGP+ICMSE, GP+IMSE, and CenGP+IMSE, respectively, with the censored regions shaded in red. The top plots of (b), (c), and (d) show the true function $\xi(\cdot)$ (black line) and the predicted function $\hat{\xi}(\cdot)$ (dashed line), with original design points (black crosses) and sequential runs (numbered). The bottom plots show the corresponding predictive standard deviation.*

2.2.3 An illustrative example

We illustrate the ICMSE criterion using a 1D example. Suppose the mean response of physical experiments is:

$$\xi(x) = 0.5 \sin\left(10(x - 1.02)^2\right) - 1.25(x - 0.75)(2x - 0.25) + 0.2, \tag{2.12}$$

with measurement noise variance $\sigma_\epsilon^2 = 0.1^2$. Further suppose censoring occurs above an upper limit of $c = 0.55$. The initial design consists of 6 equally-spaced points, which results in 5 observed points and 1 censored point. The Gaussian correlation function is used for $R_{\boldsymbol{\theta}_\xi}$, with model parameters estimated via maximum likelihood.

We compare our ICMSE method (2.11) to the standard IMSE methods (2.8), which

practitioners might use. The first method, called "GP+IMSE", uses the IMSE criterion with an *uncensored* GP model (2.6). Here, *only* the observed data $\mathbf{y}_o$ are used, since the uncensored GP model can not integrate any censored observations. The other method, called "CenGP+IMSE", uses the IMSE criterion with the *censored* GP model (2.4), which is fitted using the entire training set $\mathcal{Y}_n$. Our method is then denoted as "CenGP+ICMSE", since it uses the ICMSE criterion (2.11) with the censored GP model (2.4).

Figure 2.5(a) shows the proposed criterion for our CenGP+ICMSE method (in orange). It selects the next design point at $x_7^* = 0.068$, which balances the two desired properties from the ICMSE criterion. First, it avoids regions with high posterior probabilities of response censoring, due to the presence of $\mathbf{H}_c(\cdot)$ in (2.11). The next point x_7^*, which minimizes (2.11), subsequently *avoids* the censored regions (shaded red), as desired. In contrast, Figure 2.5(a) also shows the design criteria for GP+IMSE (green) and CenGP+IMSE (blue). We see that both IMSE methods choose the next point *within* the censored regions, as the IMSE design criterion does not consider the probability of a new observation being censored. Second, the next point x_7^* chosen by CenGP+ICMSE minimizes the overall predictive uncertainty for the mean function $\xi(\cdot)$, since the ICMSE criterion is small in regions *away* from existing design points. This can be seen within the region $[0.2, 0.5]$, where local minima of the ICMSE criterion are found between training points.

The top plots in Figure 2.5(b)-(d) show the next 3 design points (x_7^*, x_8^*, x_9^*) from the three considered design methods, as well as the final predictor $\hat{\xi}(\cdot)$ with all 9 points. The bottom plots in Figure 2.5(b)-(d) show the corresponding predictive standard deviation. We see that CenGP+ICMSE yields noticeably lower predictive uncertainty compared to the two IMSE methods, which is as desired. Table 2.1 shows the root mean-squared error (RMSE) after the 3 sequential runs over a test set of 1000 equally-spaced points. The proposed CenGP+ICMSE method achieves much smaller errors compared to the two IMSE baselines. We will provide a more comprehensive comparison of predictive performance in Section 2.3.4.

Table 2.1: *RMSE for 3 sequential runs in the 1D illustrative example* (2.12)*, using the proposed method (CenGP+ICMSE) and the two IMSE baselines (GP+IMSE, CenGP+IMSE).*

RMSE	GP+IMSE	CenGP+IMSE	CenGP+ICMSE
6 runs	0.260	0.260	0.260
7 runs	0.260	0.214	0.119
8 runs	0.260	0.236	0.102
9 runs	0.260	0.203	0.096

2.3 ICMSE design for bi-fidelity modeling

Next, we extend the ICMSE design to the bi-fidelity setting, where auxiliary computer experiment data are available. We first present the GP framework for bi-fidelity modeling, and extend the earlier ICMSE criterion. We then present an algorithmic framework for efficient implementation, and investigate its performance on two illustrative examples.

2.3.1 Modeling framework

Let $f(\mathbf{x})$ denote the *computer* experiment output at input $\mathbf{x}$. We model $f(\cdot)$ as the GP model:

$$f(\cdot) \sim \text{GP}\{\mu_f, \sigma_f^2 R_{\boldsymbol{\theta}_f}(\cdot,\cdot)\}. \tag{2.13}$$

Following Section 2.2.1, let $\xi(\mathbf{x})$ denote the latent mean response for *physical* experiments at input $\mathbf{x}$. We assume that $\xi(\cdot)$ takes the form:

$$\xi(\mathbf{x}) = f(\mathbf{x}) + \delta(\mathbf{x}), \tag{2.14}$$

where $\delta(\mathbf{x})$ is the so-called *discrepancy* function, quantifying the difference between computer and physical experiments at input $\mathbf{x}$. Following [56], we model this discrepancy using a zero-mean GP model:

$$\delta(\cdot) \sim \text{GP}\{0, \sigma_\delta^2 R_{\boldsymbol{\theta}_\delta}(\cdot,\cdot)\}, \tag{2.15}$$

where the prior on $\delta(\cdot)$ is independent of $f(\cdot)$. Here, physical experiments are observed with experimental noise as in Section 2.2.1, whereas computer experiments are observed without noise.

Suppose $(n-m)$ computer experiments and m physical experiments (n experiments in total) are conducted at inputs $\mathbf{x}_{1:n} = \{\mathbf{x}^f_{1:(n-m)}, \mathbf{x}^\xi_{1:m}\}$, yielding data $\mathbf{f} = [f_1, \cdots, f_{n-m}]$ and $\mathcal{Y}_m = \{\mathbf{y}_o, \mathbf{y}'_c \geq \mathbf{c}\}$. Note that censoring occurs only in physical experiments, since computer experiments are conducted via numerical simulations. Assuming all model parameters are known (parameter estimation is discussed later in Section 2.3.3), the mean response $\xi(\mathbf{x}_{\text{new}})$ at a new input $\mathbf{x}_{\text{new}}$ has the following conditional mean and variance:

$$\hat{\xi}(\mathbf{x}_{\text{new}}) = \mathbb{E}[\xi(\mathbf{x}_{\text{new}})|\mathbf{f}, \mathcal{Y}_m] = \mu_f + \boldsymbol{\gamma}^T_{n,\text{new}} \boldsymbol{\Gamma}_n^{-1} \left([\mathbf{f}, \mathbf{y}_o, \hat{\mathbf{y}}_c]^T - \mu_f \mathbf{1}_n \right), \quad (2.16)$$

$$s^2(\mathbf{x}_{\text{new}}) = \text{Var}[\xi(\mathbf{x}_{\text{new}})|\mathbf{f}, \mathcal{Y}_m] = \sigma_f^2 + \sigma_\delta^2 - \boldsymbol{\gamma}^T_{n,\text{new}} (\boldsymbol{\Gamma}_n^{-1} - \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\Sigma} \boldsymbol{\Gamma}_n^{-1}) \boldsymbol{\gamma}_{n,\text{new}}, \quad (2.17)$$

where $\boldsymbol{\gamma}_{n,\text{new}} = \sigma_f^2 [R_{\boldsymbol{\theta}_f}(\mathbf{x}_i, \mathbf{x}_{\text{new}})]_{i=1}^n + \sigma_\delta^2 [\mathbf{0}_{n-m}, R_{\boldsymbol{\theta}_\delta}(\mathbf{x}_i, \mathbf{x}_{\text{new}})]_{i=1}^m$ is the covariance vector, and $\boldsymbol{\Gamma}_n = \sigma_f^2 [R_{\boldsymbol{\theta}_f}(\mathbf{x}_i, \mathbf{x}_j)]_{i=1}^n{}_{j=1}^n + \text{diag}\big(\mathbf{0}_{n-m}, \sigma_\epsilon^2 \mathbf{I}_m + \sigma_\delta^2 \times [R_{\boldsymbol{\theta}_\delta}(\mathbf{x}_i, \mathbf{x}_j)]_{i=1}^m{}_{j=1}^m\big)$ is the covariance matrix. Here, $\hat{\mathbf{y}}_c = \mathbb{E}[\mathbf{y}'_c|\mathbf{f}, \mathbf{y}_o, \mathbf{y}'_c \geq \mathbf{c}]$ is the expected response for latent vector $\mathbf{y}'_c$ given data $\{\mathbf{f}, \mathcal{Y}_\mathbf{m}\}$, and $\boldsymbol{\Sigma}_c = \text{Var}[\mathbf{y}'_c|\mathbf{f}, \mathbf{y}_o, \mathbf{y}'_c \geq \mathbf{c}]$ is its conditional variance, with $\boldsymbol{\Sigma} = \text{diag}(\mathbf{0}_{n-n_c}, \boldsymbol{\Sigma}_c)$. While such equations appear quite involved, they are simply the bi-fidelity extensions of the earlier GP modeling equations (2.3) and (2.4). For simplicity, we have overloaded some notations from (2.3) and (2.4) here; the difference should be clear from the context.

2.3.2 Bi-fidelity design criterion

Now, we extend the ICMSE design to the bi-fidelity setting. The goal is to design *physical* experiment runs (which may be censored), given auxiliary computer experiment data (which are not censored). This setting is often encountered in practice, particularly in designing physical experiments for *validating* computer codes.

Under the above bi-fidelity GP model, the following theorem gives an explicit expression for the ICMSE design criterion.

Theorem 4. *With experimental data* $\{\mathbf{f}, \mathcal{Y}_m\}$*, the proposed ICMSE criterion has the following explicit expression:*

$$\begin{aligned}\text{ICMSE}(\mathbf{x}_{n+1}) &= \int_{[0,1]^p} \mathbb{E}_{Y_{n+1}|\mathbf{f},\mathcal{Y}_m} \left[\text{Var}(\xi(\mathbf{x}_{\text{new}})|\mathbf{f}, \mathcal{Y}_m, Y_{n+1})\right] \, d\mathbf{x}_{\text{new}} \\ &= \int_{[0,1]^p} \sigma^2_{\text{new}} - \boldsymbol{\gamma}^T_{n+1,\text{new}} \boldsymbol{\Gamma}^{-1}_{n+1} \mathbf{H}_c(\mathbf{x}_{n+1}) \boldsymbol{\Gamma}^{-1}_{n+1} \boldsymbol{\gamma}_{n+1,\text{new}} \, d\mathbf{x}_{\text{new}}, \end{aligned} \tag{2.18}$$

where $\sigma^2_{\text{new}} = \text{Var}[\xi(\mathbf{x}_{\text{new}})|\mathbf{f}, \mathcal{Y}_m]$*, and* $\boldsymbol{\gamma}_{n+1,\text{new}}$ *and* $\boldsymbol{\Gamma}_{n+1}$ *follow from* (2.16) *and* (2.17)*. The matrix* $\mathbf{H}_c(\mathbf{x}_{n+1})$ *has an easy-to-evaluate expression given in Appendix B.2.1.*

The proof can be found in Appendix B.2.1. The following corollary gives a simplification of (2.18) under a product correlation structure.

Corollary 1. *Suppose* $R_{\boldsymbol{\theta}_f}(\cdot,\cdot)$ *and* $R_{\boldsymbol{\theta}_\delta}(\cdot,\cdot)$ *are product correlation functions:*

$$R_{\boldsymbol{\theta}_f}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} R^{(l)}_{\boldsymbol{\theta}_f}(x_l, x'_l), \quad R_{\boldsymbol{\theta}_\delta}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} R^{(l)}_{\boldsymbol{\theta}_\delta}(x_l, x'_l), \tag{2.19}$$

with $\mathbf{x} = [x_1, \cdots, x_p]^T$*. Then, the ICMSE criterion* (2.18) *can be further simplified as:*

$$\text{ICMSE}(\mathbf{x}_{n+1}) = \bar{\sigma}^2 - \text{tr}\left(\boldsymbol{\Gamma}^{-1}_{n+1} \mathbf{H}_c(\mathbf{x}_{n+1}) \boldsymbol{\Gamma}^{-1}_{n+1} \boldsymbol{\Lambda}\right), \tag{2.20}$$

where $\bar{\sigma}^2 = \int \sigma^2_{\text{new}} \, d\mathbf{x}_{\text{new}}$*, and* $\boldsymbol{\Lambda}$ *is an* $(n+1) \times (n+1)$ *matrix with* $(i,j)^{th}$ *entry:*

$$\begin{aligned} \Lambda_{ij} &= \prod_{l=1}^{p} \left[\int_0^1 \zeta^{(l)}(x_{i,l}, x) \zeta^{(l)}(x_{j,l}, x) \, dx \right], \quad \text{and} \\ \zeta^{(l)}(z, x) &= R^{(l)}_{\boldsymbol{\theta}_f}(z, x) + \mathbb{1}_{\{i>(n-m)\}} R^{(l)}_{\boldsymbol{\theta}_\delta}(z, x). \end{aligned} \tag{2.21}$$

The key simplification from Corollary 1 is that it reduces the p-dimensional integral in the ICMSE criterion (2.18) to a product of 1D integrals, which are more easily computed.

Furthermore, if Gaussian correlation functions are used, these integrals can be reduced to error functions, which yield an easy-to-evaluate design criterion for ICMSE (see Appendix B.2.2 for details). Given the computational complexities of censored data, this simplification allows for efficient design optimization. Corollary 1 is motivated by the simplification of the IMSE criterion in [74]. The proof can be found in Appendix B.2.2.

The interpretation of the bi-fidelity ICMSE criterion (2.18) is analogous to that of the single-fidelity ICMSE criterion (2.11). Similar to the censoring adjustment function, the matrix $\mathbf{H}_c(\cdot)$ factors in the posterior probability of censoring over the input space, and is used to adjust the variance reduction term in the criterion. Viewed this way, the ICMSE criterion (2.18) provides the same design trade-off as before: the next design point should jointly (i) avoid censored regions by adaptively identifying such regions from data at hand, and (ii) minimize predictive uncertainty from the GP model.

2.3.3 An adaptive algorithm for sequential design

We present next an adaptive algorithm `ICMSE` for implementing the proposed ICMSE design. This algorithm applies for both the single-fidelity setting (with flag $I_{\mathrm{BF}} = 0$) in Section 2.2 and the bi-fidelity setting (with flag $I_{\mathrm{BF}} = 1$) in Section 2.3. First, an initial n_{ini}-point design is set up for initial experimentation: physical experiments for the single-fidelity setting, and computer experiments for the bi-fidelity setting. In our implementation, we used the maximum projection (MaxPro) design proposed by [25], which provides good projection properties and thereby good GP predictive performance. Next, the following two steps are performed iteratively: (i) using observed data $\{\mathbf{f}, \mathcal{Y}_m\}$, the GP model parameters are estimated using maximum likelihood, (ii) the next design point $\mathbf{x}^*_{n+1}$ is then obtained by minimizing the ICMSE criterion (equation (2.11) for the single-fidelity setting, equation (2.18) for the bi-fidelity setting), along with its corresponding response Y_{n+1}. This is then repeated until a desired number of samples is obtained.

To optimize the ICMSE criterion, we use standard numerical optimization methods

Algorithm 2 `ICMSE`(n_{ini}, n_{seq}, c, I_{BF}): Adaptive design under censoring

if $I_{\text{BF}} = 0$ **then** ▷ Single-fidelity
  Generate an n_{ini}-run initial MaxPro design $\mathbf{x}_{1:n_{\text{ini}}}$
  Collect initial data $\mathcal{Y}_{n_{\text{ini}}}$ at inputs $\mathbf{x}_{1:n_{\text{ini}}}$ from physical experiments
  Estimate model parameters $\{\mu_\xi, \sigma^2_\xi, \boldsymbol{\theta}_\xi\}$ using MLE from initial data $\mathcal{Y}_{n_{\text{ini}}}$
else ▷ Bi-fidelity
  Generate an n_{ini}-run initial MaxPro design $\mathbf{x}_{1:n_{\text{ini}}}$
  Collect initial data $\mathbf{f}$ at inputs $\mathbf{x}_{1:n_{\text{ini}}}$ from computer experiments
  Estimate model parameters $\{\mu_f, \sigma^2_f, \boldsymbol{\theta}_f\}$ using MLE from $\mathcal{Y}_{n_{\text{ini}}}$, and let $\sigma^2_\delta = 0$
for $k = n_{\text{ini}} + 1, \cdots, n_{\text{ini}} + n_{\text{seq}}$ **do** ▷ n_{seq} sequential runs
  if $I_{\text{BF}} = 0$ **then**
    Obtain new design point $\mathbf{x}^*_k$ by minimizing ICMSE criterion (2.11)
  else
    Obtain new design point $\mathbf{x}^*_k$ by minimizing ICMSE criterion (2.18)
  Perform experiment at $\mathbf{x}^*_k$ and collect response Y_k (which may be censored)
  Update model parameter estimates using new data

in the R package `nloptr` [75], in particular, the Nelder-Mead method [76]. The main computational bottleneck in optimization is evaluating moments of the truncated multivariate normal distribution for $\mathbf{H}_c(\cdot)$ (see equations (B.10) and (B.13) in Appendix B). In our implementation, these moments are efficiently computed using the R package `tmvtnorm` [77]. Appendix B.3 details further computational steps for speeding-up design optimization, involving an approximation of the expected variance term via a plug-in estimator. Similar to the standard IMSE criterion, the ICMSE criterion can be quite multi-modal. We therefore suggest performing multiple random restarts of the optimization, and taking the solution with the best objective value as the new design point.

2.3.4 Illustrative examples with adaptive design algorithm

We first illustrate the proposed algorithm `ICMSE` on a 1D bi-fidelity example. Suppose the computer simulation is given by

$$f(x) = 0.5 \sin\left(10(x - 1.02)^2\right) + 0.1, \tag{2.22}$$

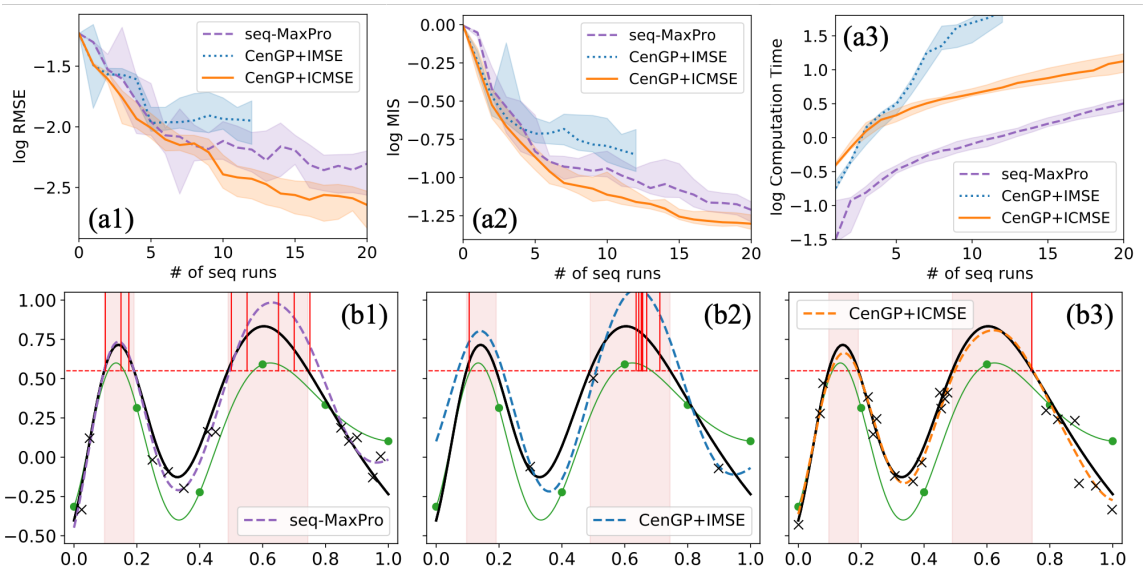

Figure 2.6: *A 1D bi-fidelity example* (2.22)*. The top plots show the log-RMSE (a1), log-MIS (a2), and log-computation time (a3, in seconds) over the number of sequential runs for each method. Solid lines mark the median over the 20 replications, and the shaded regions mark the 25%-75% quantiles. The bottom plots show the predicted functions and sequential runs (black crosses), using the three considered methods. Here, the green line marks the computer experiment $f(\cdot)$, the black line marks the mean physical experiment $\xi(\cdot)$, and the shaded regions mark the censored regions.*

with the same physical experiment settings as in Section 2.2.3. We begin with an $n_{\text{ini}} = 6$-run equally-spaced points $x^f_{1:6} = \{(i-1)/5\}_{i=1}^6$ for computer experiments. We then perform a sequential $n_{\text{seq}} = 20$-run design for physical experiments using the algorithm `ICMSE`. The Gaussian correlation function is used for both GPs. The proposed CenGP+ICMSE method is compared with the existing CenGP+IMSE method (see Section 2.2.3) and the seq-MaxPro method ("seq-MaxPro", see [78]), which provides a sequential implementation of the MaxPro design. This is then replicated 20 times.

We consider two evaluation metrics for predictive performance: RMSE and the interval score proposed in [79]. The first assesses predictive accuracy, and the second assesses uncertainty quantification. The $(1-\alpha)\%$ interval score is defined as

$$\text{IS}(\xi_l, \xi_u; \xi) = (\xi_u - \xi_l) + \frac{2}{\alpha}(\xi_l - \xi)_+ + \frac{2}{\alpha}(\xi - \xi_u)_+, \tag{2.23}$$

where $(a)_+ = \max(a, 0)$, ξ is the ground truth, and $[\xi_l, \xi_u]$ is an $(1-\alpha)\%$ predictive interval.

Here, we set $1-\alpha = 68\%$, with predictive interval $[\hat{\xi} - \sqrt{s^2}, \hat{\xi} + \sqrt{s^2}]$, where $\hat{\xi}$ and s^2 are obtained from (2.16) and (2.17). The mean interval score (MIS) is then computed over the entire test set. We also compared computation time on a 1.4 GHz Quad-Core Intel Core i5 laptop.

Figure 2.6 (a) shows the log-RMSE, log-MIS, and log-computation time for the considered methods. The proposed CenGP+ICMSE method performs constantly better over seq-MaxPro, with smaller RMSE and MIS values for most sequential runs. While CenGP+ICMSE requires more computation time compared to seq-MaxPro, it can adaptively consider the posterior probability of censoring. Here, the CenGP+IMSE method is terminated early after 12 sequential runs, due to numerical instabilities (and thereby expensive computation) in evaluating the predictive equations. This is because, by ignoring censoring, CenGP+IMSE overestimates the potential variance reduction in censored regions, leading to many sequential points very close together in such regions.

Figure 2.6 (b) shows the sequential design points and the predicted mean responses $\hat{\xi}(\cdot)$ for a single replication. Compared to the existing two methods, CenGP+ICMSE yields visually improved prediction of the true mean response $\xi(\cdot)$ in both censored and uncensored regions. One reason for this is that the ICMSE criterion chooses points which jointly (i) avoid censored regions and (ii) minimize predictive uncertainty. For (i), note that only 1/20 = 5% of sequential runs are censored for ICMSE, whereas 8/20 = 40% and 9/12 = 75% of sequential runs are censored for seq-MaxPro and CenGP+IMSE. This shows that CenGP+ICMSE effectively estimates the posterior probability of censoring, and avoids regions with high probabilities for sampling. For (ii), Figure 2.6 (a3) shows that the sequential runs from CenGP+ICMSE are far away from existing points, and also concentrated near the boundary of the censored region. Intuitively, this minimizes predictive uncertainty by ensuring design points well-explore the input space while avoiding losing information due to censoring.

Next, we conduct a 2D simulation. The computer simulation and mean physical experi-

Table 2.2: *The median RMSE, MIS, and computation time, under different sequential run sizes for the three considered design methods in a 2D bi-fidelity example* (2.24).

	RMSE			MIS			Computation Time (in s)		
Sequential runs	5	15	40	5	15	40	5	15	40
Seq-MaxPro	1.74	1.36	1.12	5.58	4.01	3.22	3.03	10.18	57.24
CenGP+IMSE	1.61	1.46	-	4.57	4.27	-	25.13	121.24	-
CenGP+ICMSE	1.40	1.21	0.97	4.58	3.80	3.01	9.77	25.01	95.67

ment functions are taken from [69]:

$$\begin{aligned} f(\mathbf{x}) =& \frac{1}{4}\xi\left(x_1+\frac{1}{20}, x_2+\frac{1}{20}\right)+\frac{1}{4}\xi\left(x_1+\frac{1}{20}, \left(x_2-\frac{1}{20}\right)_+\right) \\ &+\frac{1}{4}\xi\left(x_1-\frac{1}{20}, x_2+\frac{1}{20}\right)+\frac{1}{4}\xi\left(x_1-\frac{1}{20}, \left(x_2-\frac{1}{20}\right)_+\right), \end{aligned} \tag{2.24}$$

$$\xi(\mathbf{x}) = \left[1-\exp\left(-\frac{1}{2x_2}\right)\right]\frac{2300x_1^3+1900x_1^2+2092x_1+6}{100x_1^3+500x_1^2+4x_1+20}, \tag{2.25}$$

with measurement variance $\sigma_\epsilon^2 = 1$, and a right censoring limit of $c = 10$. We begin with an initial $n_{\text{ini}} = 12$-run MaxPro design for the computer experiment, then add $n_{\text{seq}} = 40$ sequential runs for physical experiments using `ICMSE`. This is then replicated 20 times.

Table 2.2 summarizes the median RMSE, MIS, and computation time after 5, 15, and 40 sequential runs. We see that CenGP+ICMSE yields noticeably lower RMSE and MIS, suggesting the proposed design method gives a better predictive performance. While slightly more computationally expensive than seq-MaxPro, CenGP+ICMSE is much more effective at incorporating censoring information for variance reduction, which leads to improved predictive performance.

2.4 Case studies

We now return to the two motivating applications. For the wafer manufacturing problem (which only has physical experiments), we use the single-fidelity ICMSE method in Section 2.2. For the surgical planning application (which has both computer and physical

experiments), we use the bi-fidelity ICMSE method in Section 2.3.

2.4.1 Thermal processing in wafer manufacturing

Consider first the wafer manufacturing application in Section 2.1.2, where an engineer is interested in how a wafer chip's heating performance is affected by six process input variables that control wafer thickness, rotation speed, heating laser (i.e., its moving speed, radius, and power), and heating time. The response of interest $\xi(\mathbf{x})$ is the minimum temperature over the wafer, which provides an indication of the wafer's quality after thermal processing. Standard industrial temperature sensors have a measurement limit of $c = 350°$C [80], and temperatures greater than this limit are censored in the experiment.

As mentioned earlier, certain physical experiments are not only costly (e.g., wafers and laser operation can be expensive), but also time-consuming to perform (e.g., each experiment requires a re-calibration of thermal sensors, as well as a warmup and cooldown of the laser beam). To compare the sequential performance of these methods over a large number of runs, we mimic the costly physical experiments[1] with COMSOL Multiphysics simulations (Figure 2.7(a)), which provides a realistic representation of heat diffusion physics [81]. Measurement noise is then added, following an i.i.d. zero-mean normal distribution with standard deviation $\sigma_\epsilon = 1.0°$C.

The set-up is as follows. We start with an $n_{\text{ini}} = 30$-run initial experiment, then perform $n_{\text{seq}} = 45$ sequential runs. Note that the total number of $n_{\text{ini}} + n_{\text{seq}} = 75$ runs is slightly more than the rule-of-thumb sample size of $10p$ recommended by [82] – this is to ensure good predictive accuracy under censoring. Following the earlier simulations, the proposed ICMSE method (this was "CenGP+ICMSE" in the previous section) is now compared with only seq-MaxPro, since both IMSE baselines lead to poor predictive models under censoring, and can be very time-consuming to perform (see Figure 2.5 and Table 2.2). The fitted GP models are then tested on temperature data generated (without noise) on a 200-run Sobol'

[1]The surgical planning application in Section 2.4.2 performs actual physical experiments, but provides fewer sequential runs due to the expensive nature of such experiments.

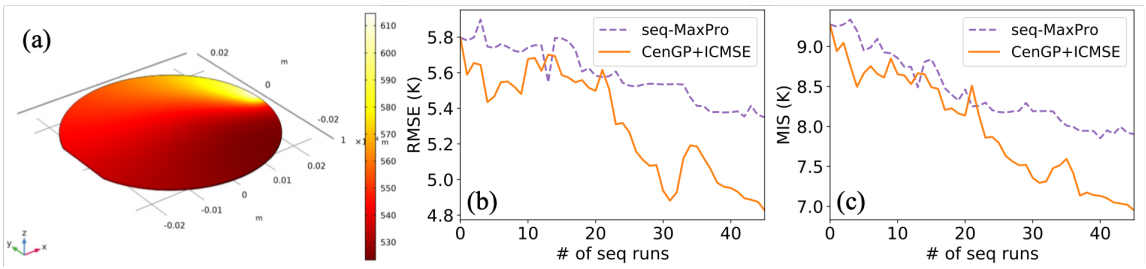

Figure 2.7: *(a) The temperature contour over the wafer chip, simulated using COMSOL Multiphysics. (b) and (c) show the RMSE and MIS of the fitted GP models over the sequential design size, respectively, for the two design methods.*

sequence [26]. Of these 200 test samples, 25 samples have minimum temperatures that exceed the censoring limit of $c = 350$°C, suggesting that roughly 12.5% of the design space leads to censoring.

Predictive performance

Figure 2.7 compares the RMSE and MIS after $n_{\text{seq}} = 45$ sequential runs. While both sequential methods provide relatively steady improvements in RMSE and MIS, the proposed ICMSE method gives a greater predictive improvement over seq-MaxPro. In particular, with 45 sequential runs, ICMSE achieves an RMSE reduction of roughly $(5.8-4.8)/5.8 = 17.2\%$ over the initial 30 runs, which is over two times greater than the RMSE reduction of $(5.8 - 5.35)/5.8 = 7.8\%$ for seq-MaxPro. A similar conclusion also holds for MIS. Despite this noticeable improvement over existing methods, there is only a moderate reduction in RMSE magnitude for ICMSE. One reason might be that the response surface (for minimal temperature over the wafer) is quite rugged and non-smooth, which makes it difficult to learn with a limited number of experimental runs.

This improved performance can again be explained by the fact that ICMSE jointly avoids censoring and minimizes predictive uncertainty. We observe that ICMSE yields no censored measurements, whereas seq-MaxPro yields 5 censored measurements (a censoring rate of $5/45 = 11.1\%$). Moreover, ICMSE *adaptively* chooses points that minimize predictive uncertainty of the GP model under censoring. This is shown in the RMSE and MIS plots in

Figures 2.7 (b) and (c): the ICMSE yields progressively lower RMSE and MIS values as sample size increases.

2.4.2 3D-printed aortic valves for surgical planning

Consider next the surgical planning application in Section 2.1.1, which uses state-of-the-art 3D printing technology to mimic biological tissues. Here, doctors are interested in predicting the stiffness of the printed organs with different metamaterial geometries. We will consider three design inputs $\mathbf{x} = (A, \omega, d)$, which parametrize a standard sinusoidal form of the substructure curve $I(t) = A\sin(\omega t)$, with diameter d (see Figure 2.2 (b) for a visualization). This parametric form has been shown to provide effective tissue-mimicking performance in prior studies [46, 24]. The response of interest $\xi(\mathbf{x})$ is the elastic modulus at a strain level of 8%, which quantifies the stiffness at a similar load situation inside the human body [46].

We use the bi-fidelity ICMSE design framework in Section 2.3, since a pre-conducted database of computer simulations is available, and we are interested in the sequential design of physical experiments. Computer simulations were performed with finite element analysis [23] using COMSOL Multiphysics. Physical experiments were performed in two steps: the aortic valves were first 3D-printed by the Connex 350 machine (Stratasys Ltd.), and then its stiffness was measured by a load cell using uniaxial tensile tests (see Figure 2.2(c); [46]). Here, physical experiments are very costly, requiring expensive material and printing costs, as well as several hours of an experimenter's time per sample. Censoring is also present in physical experiments; this happens when the force measurement of the load cell exceeds the standard limit of $15N$, corresponding to a modulus upper limit of $c = 0.23\text{MPa} = 15N(\text{force})/8mm^2(\text{area})/8\%(\text{deformation})$.

The following design set-up is used. We start with an $n_{\text{ini}} = 25$-run initial computer experiment design, and then perform $n_{\text{seq}} = 8$ sequential runs using physical experiments. The limited number of sequential runs is due to the urgent demand of the patients; in such cases, only one to two days of surgical planning can be afforded [24]. Since physical

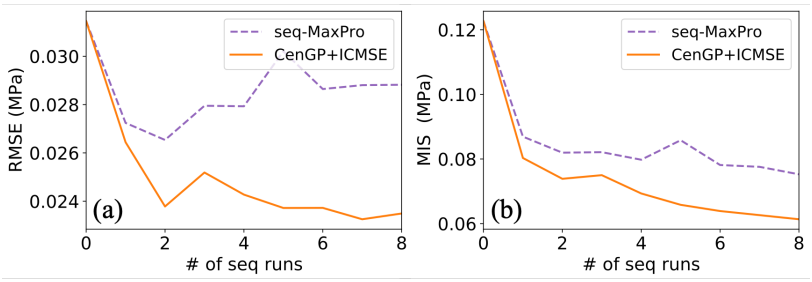

Figure 2.8: *RMSE (a) and MIS (b) for the two sequential design methods, over the number of sequential runs.*

experiments require tedious 3D printing and a tensile test (around 1.5 hours per run), this means only a handful of runs can be performed in urgent cases. As before, we compare the proposed ICMSE method with the seq-MaxPro method. The fitted GP models from both methods are tested on the physical experiment data from a 20-run Sobol' sequence. Among these 20 runs, 5 of them are censored due to the load cell limit; in such cases, we re-perform the experiment using a different testing machine with a wider measurement range. The re-experimentation is typically *not* feasible in urgent surgical scenarios, since it requires even more time-consuming tests and higher material costs.

Predictive performance

Figure 2.8 compares the predictive performance of the two design methods over $n_{\text{seq}} = 8$ sequential runs. While seq-MaxPro shows some stagnation in RMSE and MIS improvement, ICMSE yields more noticeable improvements as sample size increases. More specifically, ICMSE achieves an RMSE reduction of roughly $(0.0315 - 0.0235)/0.0315 = 25.4\%$ over the initial GP model (fitted using 25 computer experiment runs), which is around three times greater than the RMSE reduction of $(0.0315 - 0.0288)/0.0315 = 8.57\%$ for seq-MaxPro. Similar improvements can be seen by inspecting MIS. This can again be attributed to the key design trade-off. ICMSE adaptively identifies and avoids censored regions on the design space using the fitted bi-fidelity model (2.16). Here, the proposed method yields no censored measurements, whereas seq-MaxPro yields 3 censored measurements (a

Table 2.3: *RMSE on the full test set, the 5 censored runs, and the 15 observed runs, for the two sequential design methods.*

RMSE	MaxPro	ICMSE
Full	0.0288	0.0235
Censored	0.0462	0.0416
Observed	0.0199	0.0126

censoring rate of $3/8 = 37.5\%$). Furthermore, in contrast to seq-MaxPro, which encourages physical runs to be "space-filling" to the initial computer experiment runs, ICMSE instead incorporates censoring information within an adaptive design scheme, which allows for improved predictive performance.

We investigate next the predictive performance of both designs within the *censored* region. This region (corresponding to stiff valves) is important for prediction, since such valves can be used to mimic older patients [83]. We divide the test set (20 runs in total) into two categories: observed runs (15 in total) and censored runs (5 in total). The responses for the latter are obtained via new experiments on a stiffer load cell (which, as mentioned in Section 2.1.1, is typically not feasible in practice). Table 2.3 compares the RMSE of the two methods for the censored and uncensored test runs. For both methods, the RMSE for observed test runs is much smaller than that for censored test runs, which is as expected. For censored test runs, ICMSE also performs slightly better than seq-MaxPro, with $(0.0462 - 0.0416)/0.0462 = 9.9\%$ lower RMSE. One reason for this is that ICMSE encourages new runs near (but not within) censored regions (see Figure 2.6), to maximize information under censoring. Because of this adaptivity, ICMSE achieves better predictive performance within the censored region, without putting any sequential runs in this region.

Discrepancy modeling

The ICMSE method can also yield valuable insights on the discrepancy between computer simulation and reality. The learning of this discrepancy from data is important for several reasons: it allows doctors to (i) pinpoint where simulations may be unreliable, (ii) identify

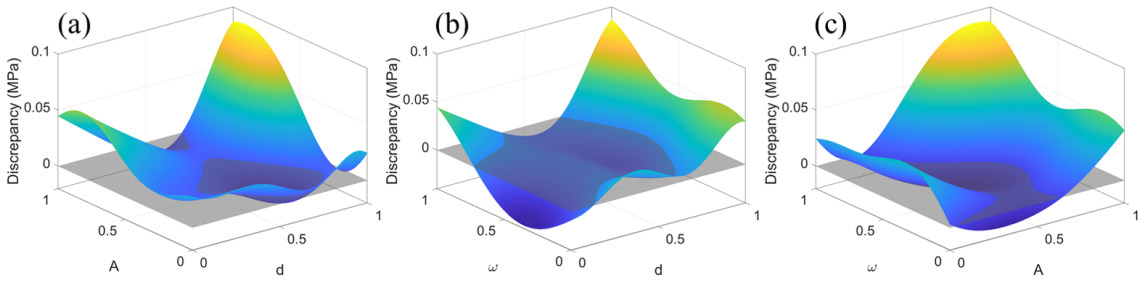

Figure 2.9: *Visualization of the estimated discrepancy $\hat{\delta}(\cdot)$ (a) over d and A, with fixed $\omega = 1$, (b) over d and ω, with fixed $A = 1$, and (c) over A and ω, with fixed $d = 1$.*

potential root causes for this discrepancy, and (iii) improve the simulation model to better mimic reality. In our modeling framework, this discrepancy can be estimated as:

$$\hat{\delta}(x) = \hat{\xi}(x) - \hat{f}(x), \tag{2.26}$$

where $\hat{\xi}(x)$ is the predictor for the physical experiment mean, fitted using 25 initial computer experiment runs and 8 physical experiment runs, and $\hat{f}(x)$ is the computer experiment model fitted using only the 25 initial runs.

Figure 2.9 shows the fitted discrepancy $\hat{\delta}(x)$ as a function of each pair of design inputs, with the third input fixed. These plots reveal several interesting insights. First, when the diameter d is moderate (i.e., $d \in [0.2, 0.7]$), Figure 2.9(a) and (b) show that the discrepancy is quite small; however, when d is small (i.e., $[0, 0.2]$) or large (i.e., $[0.7, 1]$), the discrepancy can be quite large. This is related to the limitations of finite element modeling. When diameter d is small, the simulations can be inaccurate, since the mesh size would be relatively large compared to d. When diameter d is large, simulations can again be inaccurate, due to the violation of the perfect interface assumption between the two printed polymers. Second, from Figure 2.9, model discrepancy also appears to be largest when all design inputs are large (i.e., close to 1). This suggests that simulations can be unreliable, when the stiff material is both thick ($d \approx 1$) and fluctuating ($\omega \approx 1, A \approx 1$). Finally, the model discrepancy is mostly positive over the design domain, revealing smaller stiffness evaluation via simulation compared to physical evaluation. This may be caused by the hardening of 3D-printed

samples due to exposure to natural light, as an aging property for the polymer family (e.g., see [84]). Therefore, the printed aortic valves should be stored in dark storage cells for surgical planning to minimize exposure to light.

2.5 Conclusion

In this paper, we proposed a novel integrated censored mean-squared error (ICMSE) method for adaptively designing physical experiments under response censoring. The ICMSE method iteratively performs two steps: it first estimates the posterior probability of a new observation being censored, and then selects the next design point which yields the greatest reduction in predictive uncertainty under censoring. We derived easy-to-evaluate expressions for the ICMSE design criterion in both the single-fidelity and bi-fidelity settings, and presented an adaptive design for efficient implementation. We then demonstrated the effectiveness of the proposed ICMSE method over existing methods in real-world applications on 3D-printed aortic valves for surgical planning and thermal processing in wafer manufacturing. An R package is currently in development and will be released soon.

Looking ahead, there are several interesting directions to be explored. In this work, the censoring limit c is assumed to be known. While this is true for the two motivating applications, there are other problems where c is unknown and needs to be learned from data; it would be useful to extend ICMSE for such problems. Another direction is to extend the ICMSE for experiments with truncated data. Finally, for the bi-fidelity ICMSE, it would be interesting to explore more elaborate design schemes that allow for additional computer experiments to be added sequentially.

CHAPTER 3

ACTIVE IMAGE SYNTHESIS FOR EFFICIENT LABELING

The great success achieved by deep neural networks attracts increasing attention from the manufacturing and healthcare communities. However, the limited availability of data and high costs of data collection are the major challenges for the applications in those fields. We propose in this work AISEL, an active image synthesis method for efficient labeling, to improve the performance of the small-data learning tasks. Specifically, a complementary AISEL dataset is generated, with labels actively acquired via a physics-based method to incorporate underlining physical knowledge at hand. An important component of our AISEL method is the bidirectional generative invertible network (GIN), which can extract interpretable features from the training images and generate physically meaningful virtual images. Our AISEL method then efficiently samples virtual images not only further exploits the uncertain regions but also explores the entire image space. We then discuss the interpretability of GIN both theoretically and experimentally, demonstrating clear visual improvements over the benchmarks. Finally, we demonstrate the effectiveness of our AISEL framework on aortic stenosis application, in which our method lowers the labeling cost by 90% while achieving a 15% improvement in prediction accuracy.

3.1 Introduction

Deep neural networks (NNs) [85, 86, 87] have achieved superior performance in computer vision tasks [88, 89], and attracts increasing attention from other communities, including manufacturing [90] and healthcare [91]. When fed with a *large* amount of training data (at least in the thousands [92]), NNs have shown great success in extracting high-level features and modeling complex functions. However, the available data in actual life is often *limited* and *expensive* to collect. For example, in computer-aided diagnosis of aortic stenosis, a

common yet severe heart disease [93], doctors are interested in using pre-surgical CT scans to efficiently identify the diseased patients. Here, a hospital may only have around a hundred historical records over the years, leading to unsatisfactorily performance for NNs.

In the meantime, thanks to the advances in domain research, underlining physical knowledge is often available for the learning problems in manufacturing and healthcare. Take the same aortic stenosis application as an example, the pathophysiological reason for the stenosis is mainly due to the deposited calcifications on the valve leaflets and the valve wall, and therefore change the blood flow pattern. The blood flow can be numerically simulated via computational fluid dynamics (CFD, see [94]), using the CT scans as the input geometry and boundary conditions. Incorporating such knowledge (i.e., simulation) would intuitively improve the learning model since it provides complementary information against the collected historical records. We present in this paper an *active* sampling method to incorporate underlining physical knowledge via a *complementary* dataset.

However, there are two major challenges involved in collecting such a complementary dataset. First, the inputs of the dataset (i.e., unlabeled images) are difficult to acquire in practice. This is particularly typical in the medical field, e.g., pre-surgical CT scans, due to clinical, logistic, and economic restrictions. Therefore, an effective *synthesis* model for image inputs is needed. Second, physical labeling methods are usually expensive. For example, it may take several hours of computation for a CFD model with complex geometry [94], and it would be even longer if considering the interaction of blood flow and soft biological tissue [95]. Within a practical turnaround, one can only afford a relatively *small* amount of labeled experiments. Therefore, an efficient *sampling* strategy is needed for data synthesis.

We propose in this work AISEL (an Active Image Synthesis framework for Efficient Labeling) to actively incorporate the underline physical knowledge in small-data learning. Our AISEL framework contains two major components. We first propose the generative invertible network (GIN) – a novel *bidirectional* image generative model – to encode the

actual images (i.e., the training images) into the defined lower-dimensional feature space, in which candidate virtual images can then be generated. GIN can be viewed as an extension of the generative adversarial networks [96] by adding an inverse mapping for feature extraction to the generative mapping. Moreover, we propose a new uncertainty sampling method to actively select the candidate virtual images in the GIN feature space. In our sampling method, virtual images are efficiently selected to represent the *distribution* of uncertainty in the energy-distance sense, and therefore both *exploit* the highly uncertain regions and *explore* the entire space without overlap. Labels for selected virtual images are then obtained via the physical labeling approaches at hand. By merging the training data and our AISEL dataset, improved downstream models are observed on both toy computer vision/manufacturing applications and the medical application of aortic stenosis. This paper makes the following contributions:

1. We incorporate physical knowledge into the learning process, via a complementary dataset. This ensures the incorporation of the additional information (by the *physics-based* labeling approaches), and therefore improves the downstream prediction performance.

2. We propose an efficient image sampling method for complementary dataset. Specifically, it minimizes the predictive uncertainty and mitigates the possible high labeling cost.

3. We propose a new *bidirectional* generative model – GIN for feature mapping and actively generating virtual images, conditional on the actual images. Noticeable visual improvements compared to the benchmarks are observed.

The paper is structured as follows. Section 3.2 summarizes the related works. Section 3.3.1 presents the proposed GIN with an emphasis on the difference with GAN. Section 3.4 discusses the new sampling method and features the whole AISEL learning framework. Section 3.5 demonstrates the effectiveness of our method in both toy examples and the

motivating application of aortic stenosis. Section 3.6 concludes the work with directions for future research.

3.2 Related work

Data augmentation is widely used for different learning tasks with image inputs [97, 98], via image translation, rotation and flip, and changing of the tune and/or brightness to increase the training data size. Usually, it assumes such augmentation does not change the label. However, this may not hold in, e.g., medical images. Taking CT scans as an example, different substances of human tissues correspond to different ranges of image intensity, alterations of which may lead to a completely different interpretation of the pathophysiological condition [2]. This significantly limits the augmentation methods suitable for manufacturing and healthcare applications. As to be shown later, the predictive performance with simple augmentation is not good enough.

Generative adversarial networks (GAN) [96] opens an era of adversarial training for multiple learning challenges, e.g., image segmentation [99] and domain adaptation [100]. We adapt in this work a GAN-based method for the generative model, because (i) compared to variational autoencoder [101, 102], GAN achieves visually better performance, and (ii) compared to generative flow [103], GAN contains a generative mapping from the low-dimensional features space, which can be used for our new sampling method.

To achieve efficient image sampling, study design in the feature space is desirable and crucial. However, Most GAN-based methods feature only generation mapping. Exceptions are adversarially learned inference (ALI) [104] and bidirectional GAN (BiGAN) [105], which learns both generating mapping and its inverse by a *coupled* architecture of three NNs. The model is proposed mainly for inference and representation learning. However, complicated architectures and the coupled training of three NNs requires a large amount of data, which is not suitable for our small-data learning problems. Our GIN will be compared with BiGAN to show a noticeable improvement in visual quality.

Conditional GAN (CGAN) [106] and auxiliary classifier GAN (ACGAN) [107] can generate images with given labels. Such models can be used to generate *both* virtual images and the corresponding labels for data augmentation [107, 108]. In our AISEL framework, we *only* generate the input images, while the labels are acquired via physical experiments to incorporate complementary knowledge. We will show that the proposed method has noticeable better predictive accuracy compared to the ACGAN-based method.

Transfer learning is another popular approach for small-data learning tasks [109, 110]. Adapting the models trained on natural images (mostly, ImageNet [89]), researchers are able to fine-tune the pre-trained model coefficients to address the limitation imposed by the small sample size [91]. This approach explores the visual cues extracted from natural images and assumes they are also useful in interpreting the training data at hand. However, for learning tasks in manufacturing and healthcare, the rationality of such an assumption is unclear. For example, comparing CT scans to natural images, (i) noticeable differences in image appearances are observed, and (ii) pixel intensity value has intrinsically different meanings. Nevertheless, transfer learning will be served as a baseline for the proposed framework.

Active learning (or sequential experimental design [111] in statistics literature) methods are also used for small-data learning with an oracle labeling method available [112, 113]. They aim to select the next "good" input data for labeling. Active learning methods are popular in traditional machine learning, with recent improvements for deep learning models [114, 115]. Most active learning methods in the literature assume that a sizeable *unlabeled* dataset is available. However, in manufacturing and healthcare applications, the unlabeled images are also difficult to acquire in nature.

One of the few exceptions is generative adversarial active learning (GAAL) [116] in literature, which uses GAN to generate unlabeled data. However, GAAL is proposed specifically for the support vector machine classifier. Since the support vector machine performs poorly in complicated classification tasks (e.g., our aortic stenosis application), we

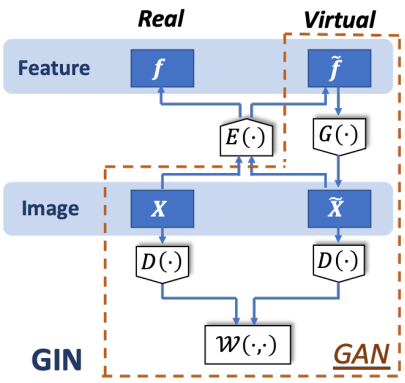

Figure 3.1: *Illustration of the proposed GIN: generator $G(\cdot)$ and discriminator $D(\cdot)$ are obtained by optimizing the Wasserstein distance $\mathcal{W}(\cdot,\cdot)$; encoder $E(\cdot)$ is a sample-to-sample inverse of $G(\cdot)$, explicitly trained by minimizing MSE. Compared to GAN, GIN contains the additional encoder $E(\cdot)$.*

will compare our method with a modified version of GAAL using a convolutional NN as the classifier.

Few-shot learning is another popular method for small-data learning tasks [117]. Though it can successfully handle learning tasks with very small training data, it usually requires many such tasks. Here, we only have one task, and therefore we will leave out few-shot learning baselines.

3.3 Generative invertible network

In this section, we propose the novel bidirectional GIN as the feature encoding and image generating model, for later efficient image sampling. We first present the GIN architecture with a detailed comparison to GAN. We then show the implementation detail and algorithm for the proposed GIN.

3.3.1 Image generating

Following the standard GAN [96, 118], we model the training set images as realizations of the distribution of the images of interest $\mathcal{X} : \mathcal{B}[\mathbb{X}] \mapsto [0, 1]$. Here, $\mathbb{X} = \mathbb{R}^{n_1 \times n_2}$ denotes the space of images with pixel size $n_1 \times n_2$, and $\mathcal{B}[\mathbb{X}]$ is its Borel set [119]. Furthermore, in order to efficiently learn the generative mapping and later interpretation, we define a

feature space $\mathbb{F} = [-1, 1]^r$. Here, r is the pre-defined dimension of the feature space, usually assumed to be much lower than that of the image space. We set a non-informative, uniform measure $\mathcal{U}$ on the feature space $\mathbb{F}$, which represents the lack of understanding of the feature space. The goal is to learn a generative mapping $G(\cdot) : \mathbb{F} \mapsto \mathbb{X}$ which best pushforwards the uniform measure $\mathcal{X}' = G_{\#}(\mathcal{U})$ and mimics the target measure $\mathcal{X}$. We use in this work the Wasserstein-1 metric [120] as the loss function:

$$\mathcal{W}(\mathcal{X}, \mathcal{X}') = \inf_{\gamma} \int_{\mathbb{X}\times\mathbb{X}} \|\mathbf{X} - \mathbf{X}'\|_2 d\gamma(\mathbf{X}, \mathbf{X}'), \tag{3.1}$$

where $\|\cdot\|_2$ is the l_2 norm, and the infimum is obtained with respect to all the possible joint distribution $\gamma : \mathcal{B}[\mathbb{X}\times\mathbb{X}] \mapsto [0, 1]$ whose marginals are $\mathcal{X}$ and $\mathcal{X}'$. We adopt the Kantorovich-Rubinstein dual form [120] of Wasserstein distance for efficient implementation:

$$\mathcal{W}(\mathcal{X}, \mathcal{X}') = \sup_{\|D(\cdot)\|_L \leq 1} \mathbb{E}_{x\sim\mathcal{X}}[D(x)] - \mathbb{E}_{x'\sim\mathcal{X}'}[D(x')]. \tag{3.2}$$

Here $D(\cdot) : \mathbb{X} \mapsto \mathbb{R}$ is an evaluating function and $\|D(\cdot)\|_L \leq 1$ represents that $D(\cdot)$ is Lipschitz-1 continuous [118].

We use a NN to approximate the generating mapping $G(\cdot)$, named *generator*, and another NN for the evaluating function $D(\cdot)$, named *discriminator*. The aim is to find the optimum of the following minimax function:

$$\min_{G(\cdot)} \max_{D(\cdot)} \mathbb{E}_{x\sim\mathcal{X}_n}[D(x)] - \mathbb{E}_{u\sim\mathcal{U}}[D(G(u))], \tag{3.3}$$

where $\mathcal{X}_n$ is the empirical measure for the training images with size n, and $\mathcal{U}$ is the uniform measure to be pushforwarded. Iterative training strategy can be adapted. Further discussion on the numerical implementation and the convergence analysis can be found in Section 3.3.3.

3.3.2 Feature encoding

Assume for now the generating mapping $G(\cdot)$ is known. We are interested in finding an encoding mapping $E(\cdot) : \mathbb{X} \mapsto \mathbb{F}$ to embed the images back to the feature space, which, to be shown in Theorem 5, is an inverse of $G(\cdot)$. Similar to the generating mapping, we use a NN to parametrize $E(\cdot)$, named *encoder*. Since the task here is to extract the feature vectors from images, a convolutional neural network (CNN) is used with mean square error (MSE) loss:

$$\min_{E(\cdot)} \mathbb{E}_{\mathbf{X}\sim\mathcal{X}_n} \|E(\mathbf{X}) - f\|_2^2, \tag{3.4}$$

where $f \in \mathbb{F}$ is the corresponding feature vector associated with the image $\mathbf{X}$. The reason we use an MSE loss is due to the desired regression task here: we want a strong metric to ensure the sample-to-sample inverse of $G(\cdot)$. Furthermore, we want $E(\cdot)$ dedicated only on this inversion task, and therefore permits an efficient sampling method later in Section 3.4.1.

However, the difficulty is that the feature f for the actual image $\mathbf{X}$ is unknown. In other words, $E(\cdot)$ cannot be learned from the dataset of actual images at hand. Therefore, we revise (3.4) to

$$E(\cdot) = \operatorname*{argmin}_{E(\cdot)} \mathbb{E}_{u\sim\mathcal{U}} \|E(G(u)) - u\|_2^2, \tag{3.5}$$

where $\mathbf{X}' = G(u)$ is the generated virtual images. Another advantage of using the virtual data points is that the data size of the virtual images can be large, since one may generate as many virtual images $\mathbf{X}'$ as needed. We expect the encoder $E(\cdot)$ learned via (3.5) using *virtual* data points (instead of the *actual* images in training set) is still the inverse of the generator $G(\cdot)$. Formally, we have the following Theorem:

Theorem 5. *Denote the target distribution measuring as $\mathcal{X}$ on image space $\{\mathbb{X}, \mathcal{B}[\mathbb{X}]\}$. Assume the generator $G(\cdot)$ is obtained by* (3.3) *with the training error $< \epsilon$ and encoder $E(\cdot)$ is obtained by* (3.5) *with the training error $< \delta$. If both $G(\cdot)$ and $E(\cdot)$ are Lipschitz-L continues, then the reconstruction error $\mathbb{E}_{x\sim\mathcal{X}}[G(E(x)) - x]^2$ can be bounded by $(L^2 + L + 1)\epsilon + L\delta$.*

This means the obtained $G(\cdot)$ and $E(\cdot)$ are inverses of each other in the sense of minimizing the reconstruction error. The proof of this theorem can be found in Appendix C.1.

The reason for introducing the encoding mapping $E(\cdot)$ as the inverse of generating mapping $G(\cdot)$ is twofold. First, we can use $E(\cdot)$ to encode the actual images as vectors in the feature space $\mathbb{F}$. They can then be used as lighthouses in $\mathbb{F}$, and provide intuitive understanding of the feature space (we will discuss this later in Section 3.5.2). Second and perhaps more important, in the following sampling method, we want to sample virtual images for better predictive performance with a limited labeling budget. In our AISEL method (see Section 3.4.1), the sampling is performed in $\mathbb{F}$ rather than the image space $\mathbb{X}$, for its lower dimension and the physical meaning. Moreover, while sampling virtual images, we need guidance from the features of the actual images. For example, one may not want to sample images that are too similar to any of the actual images to better explore the whole $\mathbb{F}$. This can be achieved by introducing a separating distance between virtual images and actual images (we will come back to this in Section 3.4.1); this needs to encode the actual images to the feature space by $E(\cdot)$.

3.3.3 Summary and algorithm for GIN

Putting everything together, the proposed GIN consists of three NNs: a generator $G(\cdot)$ for generating virtual images, an encoder $E(\cdot)$ for feature embedding, and a discriminator $D(\cdot)$ for computing the Wasserstein distance. Figure 3.1 illustrates the architecture of GIN. Note that in GIN, $G(\cdot)$ and $E(\cdot)$ is decoupled due to the limited training data. We present Algorithm 3 to train the proposed GIN. The first part of the algorithm is to train a generator $G(\cdot)$ parameterized by θ, and the second part is to train an encoder $E(\cdot)$ parameterized by γ. The generator and discriminator are coupled trained as GAN, while the additional encoder is separately trained by the virtual images sampled by $G(\cdot)$. In the small-data situation, the proposed GIN along with the associated algorithm can achieve visual improvement in practice, compared to other methods like BiGAN; we will provide a detailed discussion in

Algorithm 3 Generative invertible network

procedure GIN($\{\mathbf{X}_i\}_{i=1}^n$)
  Initialize $G_\theta(\cdot)$, $D_w(\cdot)$, and $E_\gamma(\cdot)$
  while θ has not converged **do**
    Sample $\{f_i'\}_{i=1}^m \sim \mathcal{U}$
    $L_G = -\sum_{i=1}^m D_w(G_\theta(f_i'))$
    $\theta = \theta - \alpha \nabla L_G$ ▷ Train generator
    $G(\cdot) = G_\theta(\cdot)$
    for $t = 0, \cdots, n_d$ **do**
      Sample $\{\mathbf{X}_i\}_{i=1}^m$ a batch from the actual data.
      Sample $\{f_i'\}_{i=1}^m \sim \mathcal{U}$
      $L_D = \sum_{i=1}^m D_w(\mathbf{X}_i) - \sum_{i=1}^m D_w(G_\theta(f_i'))$
      $w = w + \alpha \nabla L_D$ ▷ Train discriminator
      $w = \text{clip}(w, -\beta, \beta)$
  while γ has not converged **do**
    Sample $\{f_i'\}_{i=1}^m \sim \mathcal{U}$
    Generate $\{\mathbf{X}_i' = G(f_i')\}_{i=1}^m \sim \mathcal{X}'$
    $L_E = \sum_{i=1}^m (E_\beta(\mathbf{X}_i') - f_i')^2$
    $\gamma = \gamma - \alpha \nabla L_E$ ▷ Train encoder
    $E(\cdot) = E_\gamma(\cdot)$
  return $G(\cdot)$, $E(\cdot)$

Section 3.5.1.

One may be interested in finding out how "real" the virtual images can be generated using the proposed Algorithm 3, since multiple heuristic strategies are involved (e.g., iterative training of $D(\cdot)$ and $G(\cdot)$, and clip). Furthermore, note that the above computation is done with *samples* of actual images, i.e., the *empirical* probability measure $\mathcal{X}_n$, instead of the original probability measure $\mathcal{X}$. Therefore, we have the following theorem for asymptotic convergence.

Theorem 6. *Denote the target measure as $\mathcal{X}$ and its empirical measure represented by the training set data as $\mathcal{X}_n$. Assuming both neural networks $G(\cdot)$ and $D(\cdot)$ are obtained as the optimum of target function (3.3). Let $\mathcal{X}'$ be the measure obtained by the proposed Algorithm 3. Specifically, it is a pushforwarded measure of $\mathcal{U}$ by $G(\cdot)$, i.e., $\mathcal{X}'[S] = (G_\#(\mathcal{U}))[S] = \mathcal{U}[G^{-1}(S)]$ for any $S \in \mathcal{B}[\mathbb{X}]$. As the training data size approaches infinity, we have $\mathcal{X}' \to \mathcal{X}$ in distribution.*

The proof, following [118], can be found in Appendix C.2.

Theorem 6 suggests that, if we have enough training data, the generated images are *real* enough compared to the actual images. Specifically, it means the generated images and the measure $\mathcal{X}'$ have the following two properties. First and most importantly, the supports of the two measures are the same, i.e., $supp(\mathcal{X}') = supp(\mathcal{X})$, with probability 1.0. This is a natural corollary of Theorem 6. It means any generated virtual image $\mathbf{X}'$ can be regarded as a draw from the measure of actual images $\mathcal{X}$, i.e., $p_{\mathcal{X}}(\mathbf{X}') > 0$, where $p_{\mathcal{X}}(\cdot)$ denotes the probability density of $\mathcal{X}$. In other words, the generated images are always physically meaningful. Moreover, besides their support, the two probability measures themselves are the same asymptotically. This means the probability of generating the same group of images (e.g., CT scans of male patients, or CT scans of patients with no complications) is the same, which is an implicit requirement when endowing the feature space with physical meaning and for the following sampling method. Though in practice we are dealing with a small-data situation, it is still appealing to have this asymptotic convergence property.

3.4 AISEL Framework

We present now the proposed AISEL framework for small-data problems. For the simplicity of illustration, we assume the learning task is a *classification* problem with images inputs (this is the case of the motivating application); the proposed framework can be easily extended to regression tasks, which will not be elaborated on in this paper.

We adopt here the standard K-class classification setting, which uses input images $\mathbf{X}_i \in \mathbb{X}$ to predict the probability of assigning to each class $y_i \in [0,1]^K$. The native classifier $C(\cdot) : \mathbb{X} \mapsto [0,1]^K$, parameterized by a NN, refers to the model learned with *only* the small training data at hand. With the native model $C(\cdot)$ and GIN (i.e., generator $G(\cdot)$ and encoder $E(\cdot)$) at hand, we first propose the new sampling method to select m virtual images. We then discuss the physical labeling methods and why they are crucial in improving performance. Finally, an algorithmic framework is presented for implementation.

3.4.1 Active image sampling

We start with using the entropy [121] to quantify the uncertainty of the native model $C(\cdot)$. For any input image $\mathbf{X}_0 \in \mathbb{X}$ and the corresponding predicted label $y_0 = C(\mathbf{X}_0)$, we have

$$H(\mathbf{X}_0) = -\sum_{k=1}^{K} y_0[k] \log(y_0[k]), \tag{3.6}$$

where, $y_0 = [y_0[1], y_0[2], \cdots, y_0[K]]^T$. The reason for using entropy to quantify uncertainty can be explained as: (i) If we are sure about the class label of the input image, e.g., $y_0[1] = 1$ and $y_0[k] = 0, k = 2, \cdots, K$; the corresponding entropy is zero, meaning no uncertainty exists. (ii) Consider another extreme situation that $y_0[k] = 1/K, k = 1, \cdots, K$. One can easily check this maximizes the entropy, reflecting the maximal uncertainty for the label of that image.

The image space $\mathbb{X} = \mathbb{R}^{n_1 \times n_2}$ is too high dimensional to handle in reality. Since GIN is already obtained, we can measure the uncertainty (i.e., entropy), for any $f_0 \in \mathbb{F}$ in the *feature* space:

$$h(f_0) = H(G(f_0)) = -\sum_{k=1}^{K} E(G(f_0))[k] \log(E(G(f_0))[k]). \tag{3.7}$$

Here, we select the *features* of the complementary dataset in the feature space $\mathbb{F}$, rather than in the high-dimensional image space $\mathbb{X}$. Besides the dimensionality, our $G(\cdot)$ can capture the intrinsic structure of the image space $\mathbb{X}$ – selecting features in $\mathbb{F}$ (and then generating images via $G(\cdot)$) can ensure the existence of physical meaning. This is because any generated images $G(f_0)$ with any $f_0 \in \mathbb{F}$ is physically meaningful thanks to the $G(\cdot)$, while randomly sampled $\mathbf{X}_0 \in \mathbb{X}$ is most likely a matrix without any visual clue.

The entropy $h(\cdot) : \mathbb{F} \mapsto [0, \log K]$ can also be viewed as a (unnormalized) probability density on the measurable space $\{\mathbb{F}, \mathcal{B}[\mathbb{F}]\}$. We denote this *uncertainty measure* as μ_h.

We then propose to select the best set of m virtual images, by matching its empirical

distribution to the uncertainty measure:

$$f'_{1:m} = \underset{f'_{1:m}}{\operatorname{argmin}}\, dist(\mathcal{F}'_m, \mu_h). \tag{3.8}$$

Here, $dist(\cdot, \cdot)$ is a distance metric, $f'_{1:m} = \{f'_i\}_{i=1}^m$ denote the selected features, and $\mathcal{F}'_m$ denotes the empirical measure for $f'_{1:m}$. Intuitively, (3.8) means to assign more points to higher uncertainty regions (of the native model), and therefore exploit those regions. Furthermore, if taking the Bayesian perspective, it can be viewed as changing the initial uniform distribution, i.e., the non-informative prior, to the posterior distribution of uncertainty given the actual training dataset.

Motivated by the literature in the statistical community [122, 123], we select the energy distance as the metric $dist(\cdot, \cdot)$ between distributions. Therefore, we minimize:

$$\min_{f'_{1:m}} \sum_{i=1}^{m} \mathbb{E}_{\gamma \sim \mu_h} \|f'_i - \gamma\|_2 - \frac{1}{2m} \sum_{i=1}^{m} \sum_{j=1}^{m} \|f'_i - f'_j\|_2. \tag{3.9}$$

Note again, the above sampling optimization is conducted in the feature space. We observe from (3.9) that the selected features not only try to match the target uncertainty measure in the expectation sense (the first term), but also separate from one another (the second term). The separating property is of great importance; this is because any two selected features that are too close to each other can be viewed as a waste of the expensive labeling process.

Furthermore, the selected features should also be separated from the features of the *actual* images, again to avoid waste. This can be taken into account by the following modification of (3.9):

$$\min_{f'_{1:m}} \sum_{i=1}^{m} \mathbb{E}_{\gamma \sim \mu_h} \|f'_i - \gamma\|_2 - \sum_{i=1}^{m+n} \sum_{j=1}^{m+n} \frac{\|f'_i - f'_j\|_2}{2(m+n)}, \tag{3.10}$$

where n is the size of the actual dataset and let $f'_i = f_i$ for actual images with indies $i = m+1, \cdots, m+n$. Comparing (3.10) to (3.9), we notice the difference lies in the

second term, where the separating property is incorporated not only between the selected features but also between the selected features and the features of actual images. We use (3.10) for sampling features, and then generate an AISEL dataset via $G(\cdot)$. The following theorem ensures the generated AISEL dataset follows the target uncertainty measure in distribution.

Theorem 7. *Let target uncertainty measure be μ_H in* (3.6) *and the selected features by* (3.10) *be $f'_{1:m}$ with size m. Assume the $G(\cdot)$ is continuous. Further denote the set of virtual images $\{\mathbf{X}'_i\}_{i=1}^m = \{G(f'_i)\}_{i=1}^m$, with its empirical measure $\mathcal{X}'_m$. We then have $\mathcal{X}'_m \to \mu_H$ in distribution.*

Here, we show the convergence in the distribution of the *images* (rather than the *features*), as the images are the quantity of interest. Therefore, a continuous assumption on $G(\cdot)$ is needed according to continuous mapping theorem. The proof can be found in Appendix C.3; it follows from [123].

The proposed sampling strategy (3.10) reveals an important trade-off. Consider the first term, where the selected features are forced to be close to the target uncertainty measure. Since the density of our target measure (3.6) is high when the uncertainty is high, the selected features can be viewed as exploiting the highly uncertain regions. Now consider the second term, where the separating distance is maximized. This suggests the selected features should be away from (i) one another and (ii) the features for actual images. Therefore, selected features are forced to be spread out and fill the whole feature space – they explore the entire feature space. Putting both parts together, selected features for virtual images jointly exploit the highly uncertain regions and explore the entire image space. This trade-off of our AISEL dataset will be shown as the key to improve the classification performance.

We want to make a few remarks here. First, the proposed sampling method, specifically the uncertainty measure (3.7), is specifically for the classification problem at hand. With a different uncertainty measure, the proposed approach is also suitable for regression problems. For example, one may obtain the measure via predictive variance using kernel

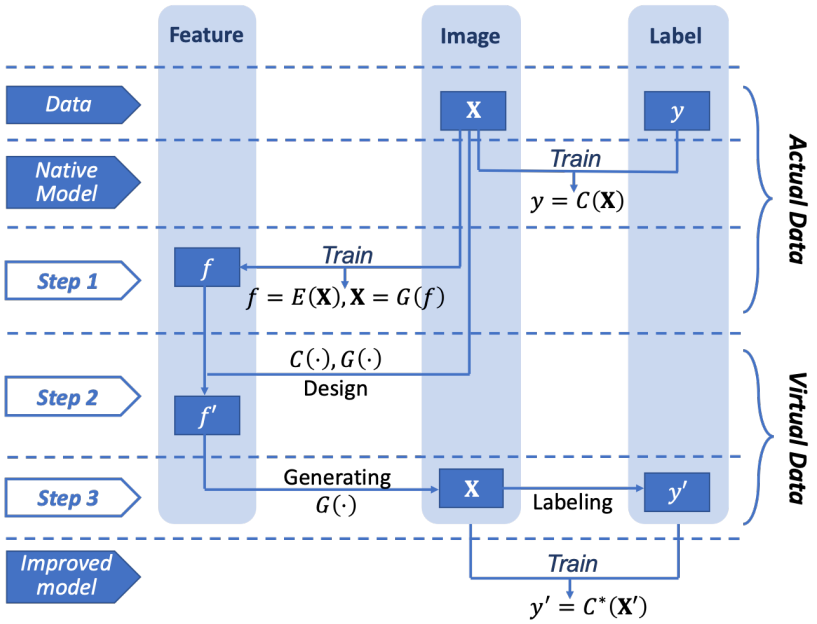

Figure 3.2: *The proposed three-step framework AISEL to efficiently sample AISEL dataset and improve classification.*

regression [124] or kriging [7] methods. Second, our method is motivated by active learning literature [112], where the next input is selected from a pool of candidates with maximal uncertainty. Different from those methods, our method (i) conducts sampling in a much lower-dimensional GIN feature space due to the intrinsic structure of image space and computational efficiency and (ii) sample a batch of images for labeling to both explore and exploit the design space. Moreover, it is worth pointing out that the proposed method is possible only when *both* the generating mapping $G(\cdot) : \mathbb{F} \mapsto \mathbb{X}$ and the encoding mapping $E(\cdot) : \mathbb{X} \mapsto \mathbb{F}$ are available via GIN. In particular, $G(\cdot)$ is used to generating images based on the selected features, while $E(\cdot)$ is used to embed the features of the actual images to guide the sampling. That is the key reason why we propose a *bidirectional* GIN in Section 3.3.1. Last but not least, the proposed method can also be used to balance the label distribution with a modification of our uncertainty measure (3.7). See Appendix C.6 for more discussion.

3.4.2 Labeling by physical principles

A key component of the proposed AISEL framework, different from data augmentation, is the incorporation of *physical knowledge* while learning. This is due to the circumstances of real-world applications in manufacturing and healthcare: (i) the size of the historical records is small, leading to a poor learning model; and (ii) thanks to the advances in domain research, physical knowledge is oftentimes available yet expensive in implementation. Therefore, we want to build a bridge to efficiently combine both the historical data (via the learning model) and physical knowledge (via physical labeling). The resulting model can be viewed as one that has learned from data and been taught by physical knowledge, and therefore better performance can be achieved.

Here, we *efficiently* incorporate physical knowledge via a complementary AISEL dataset. Specifically, we separately acquire the input image and the output label. For virtual images, (3.10) is used to efficiently sample a set of features to minimize the predictive uncertainty, and GIN is then used to map those features to images. Meanwhile, for the labels, we use the physical labeling method at hand. We then combine the actual dataset and AISEL dataset to learn the downstream classification model. With the proposed uncertainty sampling method, our AISEL dataset (i) contains complementary information from physical knowledge, and (ii) efficiently exploit and explore the image space, and therefore improves the downstream learning performance.

Lots of different physics-based labeling approaches are available. For example, finite element analysis [23] and computational fluid dynamics [94] can solve partial differential equations (i.e., representation of physical knowledge) numerically. These methods can be used to label the input images in, e.g., manufacturing applications. Physical experiments can also be applied. With the advances in additive manufacturing, tissue-mimicking 3D printing [24] with an in-vitro study [3] can be used for medicine-related learning tasks. If none of the above exists, one can also consult the experts or use certain empirical physical relationships; these methods can be used in computer vision problems. The specific approach should be

Algorithm 4 Improving classification by AISEL framework

Native model
Train CNN, $C(\cdot)$ = CNN($\{\mathbf{X}_i, y_i\}_{i=1}^n$)
Step 1: Train GIN
Set the feature space $\mathbb{F} = [-1, 1]^r$
Set prior uniform measure on $\mathbb{F}$
Train $G(\cdot)$,$E(\cdot)$ = GIN($\{\mathbf{X}_i\}_{i=1}^n$) by Algorithm 3
Step 2: Sampling features
Obtain the uncertainty measure μ_h by (3.7)
Obtain features for actual images, $f_i = E(\mathbf{X}_i)$
Optimize (3.10) by CCP and obtain features f_j'
Step 3: Acquiring AISEL dataset
Generate actual images, $\mathbf{X}_j' = G(f_j')$, $j = 1, 2, ..., m$
Obtain labels, y_j' of $\mathbf{X}_j'$ by physical approaches
Improved model
Train CNN, $C^*(\cdot)$ =CNN($\{[\mathbf{X}_i, \mathbf{X}_j'], [y_i, y_j']\}$)

made on a case-by-case basis, with the available resources at hand.

It is important to note that labeling one input image via physical-based approaches is usually *expensive*. For example, in medicine-related applications, it may take several hours of computation for a CFD model with complex geometry [94], and it would be even longer if considering the interaction of blood flow and soft biological tissue [95]. This is one of the reasons for introducing an efficient and effective sampling method to design our AISEL dataset. Viewed this way, our approach can also be used to address the problem where an expensive simulator available, and we want to use that simulator actively for the classification tasks.

3.4.3 Summary of the AISEL framework

In summary, we propose an AISEL framework to efficiently incorporate physical knowledge at hand and improve the classification performance. The native model $C(\cdot)$ can be first obtained using the small training data. As illustrated in Figure 3.2, our AISEL framework contains three steps. First, the proposed GIN is trained using the actual images, providing a feature space $\mathbb{F}$, and bidirectional mappings between it and the image space (i.e., the

generating mapping $G(\cdot)$ and the encoding mapping $E(\cdot)$). Second, the uncertainty of $C(\cdot)$ at different locations in $\mathbb{F}$ is quantified via entropy, and then the features for virtual images are sampled via (3.10). Third, virtual images are generated by $G(\cdot)$, and then labeled by the physics-based approach. Finally, the additional AISEL dataset is merged to the original training set, and an improved classifier $C^*(\cdot)$ can be trained. With the proposed AISEL framework to actively incorporate complementary knowledge via labeling, we will show later in the experiments, $C^*(\cdot)$ can achieve better classification accuracy.

We propose Algorithm 4 for our AISEL framework. In our implementation, the native model and improved model are parameterized by CNN, for its popularity in the image classification tasks. Other, perhaps more advanced architecture (e.g., ResNet [125]) can also be used. The optimization of (3.10) can efficiently implement by the convex-concave procedure (CCP, see [123]). Note that for all the NNs, especially the native model and the GIN, data augmentation methods (e.g., rotation and horizontal flip) are used. Note that our method can also be used for sequential implementation – run Algorithm 4 interactively to generate a series of AISEL datasets and therefore provide even better improvement, if budget allows.

3.5 Experiments

We first conduct toy computer vision experiments, and provide more insights on our AISEL framework. We then deploy the proposed method to the medical application of aortic stenosis, with emphasis on the pathophysiological meaning of the proposed framework.

3.5.1 Toy computer vision applications

We conduct experiments on small versions (400 in total for training) of two single-channel (i.e., grayscale) computer vision datasets – Fashion [126] and MNIST [127]. The two datasets are of particular interest due to their visual similarity to the images in the manufacturing process and modeling. For example, the images captured by a thermal camera

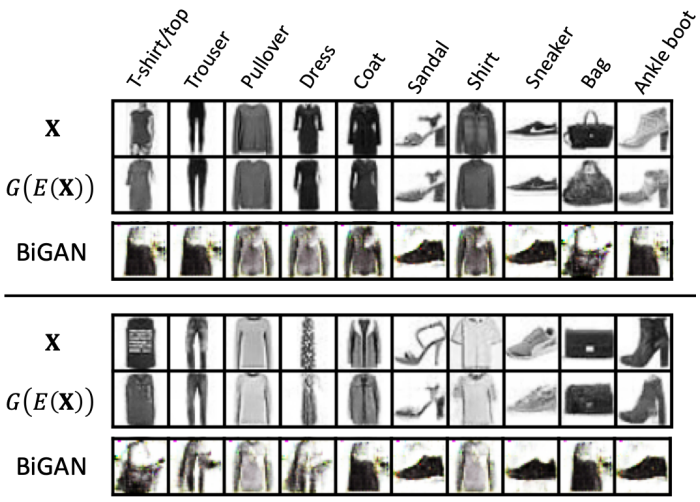

Figure 3.3: *Qualitative results for our GIN on Fashion data, including the training data* $\mathbf{X}$ *of all ten classes, our reconstructions* $G(E(\mathbf{X}))$ *and reconstructions via BiGAN [105].*

(or simulated via finite element analysis), representing a gray-level temperature contour, can be used to predict the throughput in steel manufacture and conduct quality control in semiconductor manufacturing [128]. In the following subsections, we illustrate the visual performance of the proposed GIN, and the improvement in classification by our AISEL method, *only* on the Fashion dataset; similar observation also applies for MNIST (see Appendix C.5).

Fashion dataset

The Fashion dataset [126] is an MNIST-like dataset of Zalando's article images. As shown in Figure 3.3 (see the rows $\mathbf{X}$), it contains ten classes of outfits. We observe that the images associated with classes "T-shirt/top", "coat" and "shirt" are visually similar in nature, resulting in a more challenging classification task than MNIST. Lots of works have been dedicated to classifying the Fashion dataset [129], and the leading accuracy is 96.7% by WideResNets [130]. We use this model as our labeling approach (see Section 3.5.1 for a detailed discussion).

The original Fashion dataset contains a large amount of training data (60,000 in total). To mimic the real small-data situation in manufacturing, we randomly sample 400 as our

Table 3.1: *The classification accuracy applying our AISEL method and baselines, on the Fashion dataset and MNIST.*

	Native (400)	Transfer	ACGAN	Rand (+400)	Rand (+5000)	**AISEL** (+400)	AL (+400)	Oracle (800)	Oracle (all)
Fashion	**72.8%**	74.3%	72.3%	76.4%	80.7%	**81.9%**	78.2%	81.3%	96.7%
MNIST	**88.2%**	87.4%	85.9%	90.2%	91.6%	**91.2%**	90.4%	91.3%	99.2%

training set, with roughly 40 data per label. The original test set (10,000 in total) remains untouched.

Visual results of GIN

We test first the proposed GIN. The dimension of the feature space is set to be $r = 2$, i.e., $\mathbb{F} = [-1, 1]^2$. This is only for visualization purposes; for actual employment (and in the later application of aortic stenosis), we suggest using a higher r for better performance. The detailed architecture of the three NNs can be found in Appendix C.4. Figure 3.3 shows the generated images (see the rows $G(E(\mathbf{X}))$). Visually, they look sharp and reasonable without apparent mode dropping.

Reconstruction test. To visually show the encoder $E(\cdot)$ is the inverse of generator $G(\cdot)$, we conduct the following reconstruction test: for any actual image $\mathbf{X}_i$, its feature is extracted $f_i = E(\mathbf{X}_i)$, and then a reconstructed image is generated based on that feature $G(f_i) = G(E(\mathbf{X}_i))$, which is compared with the actual $\mathbf{X}_i$ visually. The test results of all ten classes are shown in Figure 3.3, with $\mathbf{X}$ denoting the actual images, and $G(E(\mathbf{X}))$ denoting the reconstructing ones. The similarity between the two is noticeable, especially in the sense of the same class. Note that we have already proven that $G(\cdot)$ and $E(\cdot)$ are inverses of each other in an ideal situation (see Theorem 5), and here we show the inverse can be achieved in practice.

Comparison with BiGAN. We compare our GIN with BiGAN in literature, which also features a bidirectional mapping [105]. The architecture of BiGAN is set to be similar to GIN, with the same feature dimension $r = 2$ and hidden layers. Figure 3.3 (see the rows

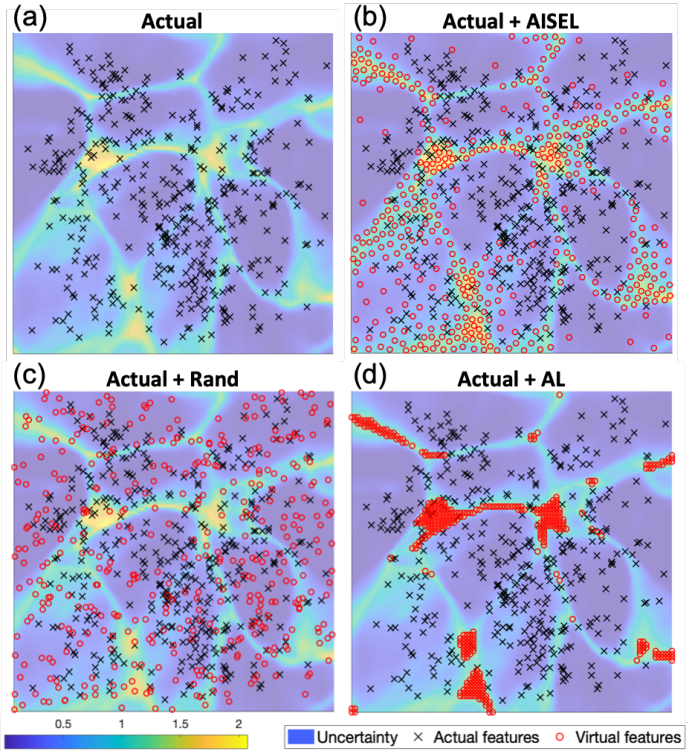

Figure 3.4: *A comparison of the selected features by our AISEL method, the random sampling method and the active learning method, with uncertainty measure* (3.7) *as background.*

"BiGAN") shows the reconstruction test conducted by BiGAN using the same set of actual images $\mathbf{X}$ as GIN. Further tuning (e.g., learning rate and hidden dimensions) of the BiGAN is also conducted, with similar performance (also see Figure 4 in [105]). In contrast, our GIN is easier to tune, and more importantly, achieves noticeable improved visual performance – better reconstruction results with no mode dropping even in this small-data situation. The reason for this difference contributes to the essentially different objectives of the two methods. BiGAN uses one discriminator to supervise both the generator and the encoder for representation learning purpose or even latent regression [104]. On the contrary, the objective of our GIN is to find the best inverse mapping for efficient sampling. Therefore, sequential order of training $G(\cdot)$ and $E(\cdot)$ is implemented in the proposed GIN to ensure the sample-to-sample inverse is *explicitly* trained by MSE metric. Therefore, we leave out the comparison of BiGAN for downstream classification.

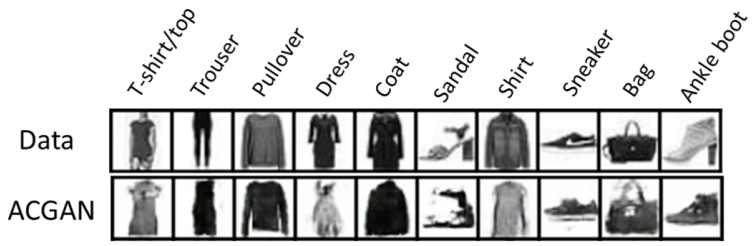

Figure 3.5: *Qualitative results of generated images of all ten classes via ACGAN baseline.*

AISEL framework

Now we test the rest of our AISEL framework. The native model $C(\cdot)$ is set to be a CNN with detailed architecture specified in Appendix C.4. The classification accuracy of the native model is only 72.8%, since only 400 data are used as the training set. We then generate an AISEL dataset with size 400. Note that the labels are obtained by the oracle model (using all 60,000 training data and WideResNet architecture, denoted as "Oracle (all)" in Table 3.1), mimicking the process of labeling by a domain expert. An improved classification model $C^*(\cdot)$ can then be obtained by the actual data and AISEL data. Final classification accuracy on the same test set is 81.9%, an almost 10% increase. This improvement shows that the proposed AISEL framework can indeed improve the predictive accuracy in the classification tasks. The reasons are that (i) the additional knowledge, i.e., labeling by the oracle model, is incorporated in the learning process, and (ii) the proposed sampling method ensures that our AISEL dataset explores the feature space. The latter will be discussed in detail below, with comparison to different baseline methods. Table 3.1 shows the final classification performance of the proposed method on both Fashion and MNIST, compared to the baselines.

Features of AISEL dataset. We first visualize the actual features (i.e., the embedded features of the actual images) in the feature space $\mathbb{F} = [-1, 1]^2$. Figure 3.4 (a) shows the 400 actual features (in black crosses) on $\mathbb{F}$, with the background visualizing the uncertainty measure (3.7). Specifically, yellow regions indicate high uncertainly of the native model $C(\cdot)$. Figure 3.4 (b) shows the features for 400 AISEL dataset (in red circles) together with

the actual features. We observed the key trade-off as mentioned: our AISEL features jointly (i) exploits the highly uncertain regions and (ii) explores the whole feature space. On one hand, objective (i) is achieved by sampling more points in the regions where uncertainty is high, i.e., those with a yellow background. Visually, the AISEL features approximately follow the uncertainty measure. On the other hand, objective (ii) is achieved by spreading the AISEL features over the whole feature space with no features too close to one another. Visually, there are no big "holes", and no two points overlap. Therefore, our AISEL method achieves a high (81.9%) classification accuracy.

Comparison with GIN-based random sampling. We compare the proposed sampling method to random sampling. Specifically, we uniformly sample 400 features, generate virtual images using those features, and then label them using the same oracle model (i.e., Oracle (all)). Figure 3.4 (c) shows the randomly generated features. We see that those features are (i) not exploiting (i.e., placing more points in) the highly uncertain regions, and (ii) overlapping with one another and to the actual features, leading to poor exploration. Therefore, the classification accuracy of the 400 randomly sampled virtual images is only 76.4%, which is noticeably lower ($81.9\% - 72.8\% = 9.1\%$) than the proposed AISEL method. If increasing the number of random virtual images to 5000, the classification accuracy can be increased to 80.7%. Our AISEL method achieves slightly higher accuracy ($81.9\% - 80.7\% = 1.2\%$), however much less virtual images, and therefore much lower labeling cost.

Comparison with active learning. We compare the proposed sampling method to an active learning (AL) method. Specifically, we adapt a similar setting in GAAL [116] in literature, using the GIN to generate a potential unlabeled dataset. To sample a virtual dataset, we set a grid (with size 101×101) on the feature space, and then select the top 400 features among the grid, whose uncertainty is the highest. Virtual images are then generated and labeled by the oracle model. Figure 3.4 (d) shows the features selected by our setting of AL. We observe that, though the selected features locate in the highly uncertain regions, they

are too close to one another. Furthermore, the selected features do not explore the whole feature space. The final classification performance of this AL is 78.2%, which is slightly better than the random sampling ($78.2\% - 76.4\% = 1.8\%$). However, our AISEL method, jointly explore and exploit the feature space, achieves a better classification performance ($81.9\% - 78.2\% = 3.7\%$).

Comparison with ACGAN. Another approach also for small-data tasks is the ACGAN-based data augmentation method [108], which generates a set of images based on the chosen labels. We train an ACGAN [107] using the actual dataset at hand. The complexity of the ACGAN is set to be similar to our GIN. Figure 3.5 visualizes the generated images of all ten classes. We see that due to the limited training size (400 in total), the images can sometimes be wrongly labeled – in Figure 3.5, the generated "Trouser" is visually more close to "Dress". The final classification accuracy is 72.3%, i.e., it offers similar classification accuracy as the native model. This is because adapting ACGAN, the labels of the augmented dataset are obtained by the training set *without* complementary information. The data size is increased, however, those labels may not be accurate; therefore, little improvement is observed in this experiment. In our AISEL framework, we use an additional labeling method (i.e., by the oracle model with an accuracy of 96.7%) for more precise labels. Therefore, our method achieves better predictive performance.

Comparison with transfer learning. Transfer learning is also popular for small-data tasks. In our setting of transfer learning, we adapt a pre-trained ResNet18 [125] (by ImageNet [89]) and only fine-tune the last fully connected layer using the training data (400 in total) at hand. The classification accuracy of the transfer learning is 74.3%, only slightly better than the native model with an accuracy of 72.8%. The reason for this is that images of the Fashion dataset are virtual different from the natural images in the ImageNet. This observation is typical in the applications of manufacturing and healthcare, where the input images are, e.g., images from a thermal camera or flow velocity contour. In the transfer learning setting [131], we implicitly assume the parameters learned by ImageNet data can use be used to

interpret the current Fashion dataset at hand. From the result of classification accuracy, the above assumption may not be valid. Our AISEL framework incorporates more accurate knowledge from physical experiments or experts, and therefore better classification model can be obtained.

Comparison with native model using 800 actual data. Another interesting baseline, though not feasible in real applications, is directly using 800 *actual* data to train an oracle model. In our setting, the first 400 data is the same as the 400 for the native model, and the remaining 400 is again randomly selected from the actual training set of the Fashion dataset. The same architecture as the improved model $C^*(\cdot)$ is used. We observe the classification accuracy is 81.3% (denoted as "oracle (800)" in Table 3.1), which is similar to our AISEL method with accuracy 81.9%. This is again due to our efficient sampling method, which both explores and exploits the image space. Meanwhile, this also verifies the good generating performance of our GIN.

3.5.2 Aortic stenosis application

We now go back to the motivating application of aortic stenosis (AS). An anonymous image dataset containing 168 patients with aortic stenosis is collected (by Piedmont Healthcare, Atlanta). For each patient, pre-surgical CT scans and the corresponding calcification amount are acquired. The learning task is to classify the calcification level as high or low, which is an important yet challenging clinical problem. Four-fold cross validation strategy is used (see Appendix C.4), leading to only 126 data as the training set. We first provide more background information on the medical problem and our dataset. We then visualize the GIN performance, with a focus on the pathophysiological meaning. Finally, we discuss the classification accuracy, compared with baselines.

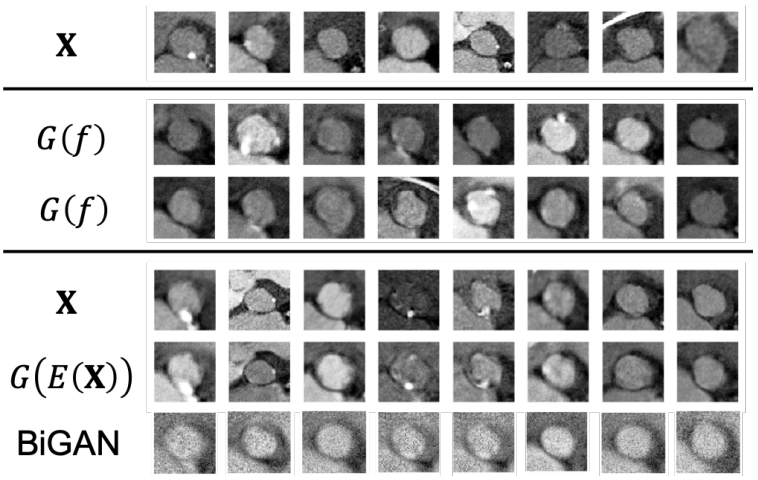

Figure 3.6: *Qualitative GIN results for the aortic stenosis application, including actual data* $\mathbf{X}$*, generated samples* $G(f)$*, and corresponding reconstructions* $G(E(\mathbf{X}))$*.*

Background on aortic stenosis

Aortic stenosis (AS) is one of the most common and most serious valvular heart diseases. Transcatheter aortic valve replacement (TAVR) is a less-invasive treatment option for severe AS patients who are at high risk for open-heart surgery. One of the major post-procedural complications of TAVR is the paravalvular leakage (PVL), i.e., blood flow leakage around the implanted artificial valve due to incomplete sealing between the implant and the native aortic valve, which is often caused by the *calcifications* presented at the aortic annulus region (a ring-shaped anatomic structure connecting the left ventricle and the aortic valve). Therefore, in clinical practice, the amount and the distribution of annular calcifications are of great importance to predicting the occurrence of post-TAVR PVL. However, in-vitro study [3] is quite costly, requiring expensive operation costs of CT scanner, as well as several days of an experimenter's time per virtual patient. Because of this, we simplify the task of PVL prediction to the task of calcification evaluation, which is deemed as an important clinical indicator of PVL risk. Due to the variant contrast level in the aortic root and the fast motion of the valve leaflets, it remains challenging to accurately evaluate calcification near the aortic annulus in pre-TAVR CT images.

Table 3.2: *A comparison of classification accuracy (accu., %), sensitivity (sens., %), specificity (spec., %), and F1 score (%) of the native model and different improved models in a 4-fold cross-validation, with data size included.*

	AISEL (+1134)				**Random (+10000)**			
Fold	Accu.	Sens.	Spec.	F1	Accu.	Sens.	Spec.	F1
1	76.19	73.91	78.95	74.42	78.57	78.26	78.95	77.27
2	73.81	76.19	71.43	71.43	76.19	71.43	80.95	73.17
3	80.95	76.47	84.00	76.92	85.71	82.35	88.00	82.05
4	71.43	75.00	68.18	69.77	73.81	70.00	77.27	70.00
Ave	**75.60**	75.39	75.64	73.14	**78.57**	75.51	81.29	75.62
	Native Model (126)				**Random (+1134)**			
Fold	Accu.	Sens.	Spec.	F1	Accu.	Sens.	Spec.	F1
1	64.29	52.17	78.95	61.54	69.05	60.87	78.95	68.29
2	56.96	47.26	66.67	48.65	64.29	61.90	66.67	60.00
3	57.14	50.00	63.64	52.63	71.43	52.94	84.00	58.82
4	59.52	55.00	63.64	56.41	64.29	60.00	68.18	60.00
Ave	**59.48**	51.11	68.23	54.81	**67.27**	58.93	74.45	61.78

Pathophysiological interpretability of GIN

We first visualize the performance of GIN. Here, the dimension of the feature space is $r = 20$, i.e., $\mathbb{F} = [-1, 1]^{20}$, considering the complexity of the CT scans. The detailed architecture of the GIN can be found in Appendix C.4.

Pathophysiologically-interpretable feature space. To better visualize the 20-dimensional feature space $\mathbb{F} = [-1, 1]^{20}$, Figure 3.7 shows a randomly selected 2D cross-section of $\mathbb{F}$ with the generated virtual valve images located at their projected feature locations; the full and enlarged version of the figure is shown in Figure C.3 in Appendix C.7. We notice that the variation of the virtual images on the feature grids is continuous and smooth. Furthermore, we observe that the two axes of the 2D cross-section shown in Figure 3.7 (also see Figure C.3) have pathophysiological meaning. As shown in the red box (enlarged images on the left side), the vertical axis can be interpreted as the change of the calcification (i.e., the regions of high intensity in the CT images) amount. As shown in the blue box (enlarged images on the right side), the horizontal axis can be interpreted as the change of valve shape and the

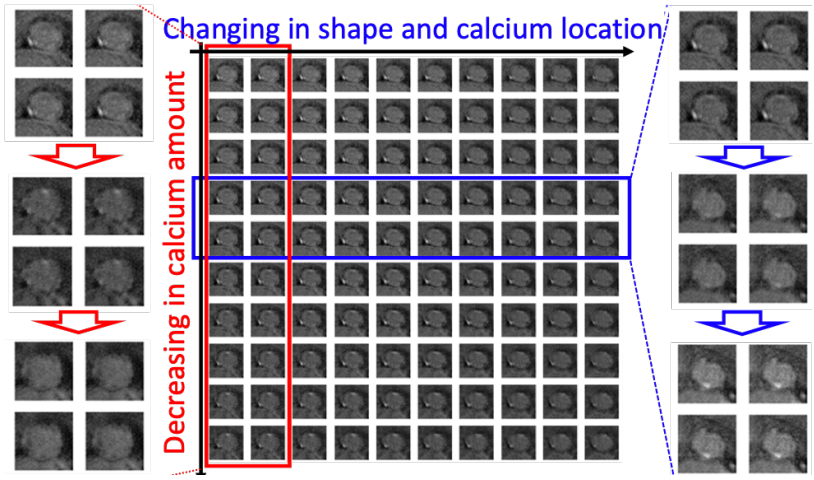

Figure 3.7: *Qualitative visualization of 2D cross-section of feature space with the generated virtual images on the (partial) grid of feature space. The full and enlarged version of the figure is shown in Figure C.3 in Appendix C.7.*

calcification location. Similar observations can be found in the other vertical or horizontal groups of images, which demonstrate the potential pathophysiological interpretability of $\mathbb{F}$.

Reconstruction test. Similar to the toy experiments, Figure 3.6 shows the reconstruction test comparing the actual CT scans $\mathbf{X}$ and the reconstructed images $G(E(\mathbf{X}))$. Visually, the reconstructed images are almost identical to the actual images with similar background color and valve geometry. Furthermore, the most important pathophysiological indicators, i.e., the location and size of the calcifications are well-recovered. This shows that: (i) using GIN can capture the features of important pathophysiological meaning, and (ii) $G(\cdot)$ and $E(\cdot)$ are inverses of each other. In Figure 3.6, we also compare the proposed GIN with BiGAN (see Section 3.5.1), with better performance observed.

Improving classification by our AISEL method

Here, we use the CT scans at the annulus to predict the calcification level (see Section 3.5.2). The native model utilizes a simple CNN structure, with the detailed architecture described in Appendix C.4. Table 3.2 summarizes the classification accuracy, sensitivity and specificity of the four-fold cross validation using the native model. For each fold, we generate an AISEL dataset with the size of 1134. In addition, two randomly sampled virtual

dataset, with the size 1134 and 10000 are also generated as comparison. We will leave out the other baselines, since a detailed comparison is already conducted in the toy experiments (see Section 3.5.1). As for labeling the virtual patients, an empirical approach is performed: a mixture model of two Gaussians is used to model the pixel intensity, based on whether the pixels are classified as normal tissues or calcifications. The volume of the calcification region is then calculated. After that, a manual check is performed by a radiologist and the calcification levels are corrected if needed. Note that if budget allows, a more sophisticated labeling approach can be used.

Classification performance. The three generated virtual datasets (proposed, random with size 1134 and 10000) are fused with the actual dataset to obtain improved classifiers. Table 3.2 summarizes the prediction accuracy together with the sensitivity, specificity, and F1 score of the different classifiers. We see that the native model performs the poorest over the test set, with less than 60% averaged accuracy. The prediction accuracy improves to 67.27% when using randomly generated samples with the size of 1134. Using our AISEL method, the averaged accuracy improves to 75.60% with the same size (126 actual + 1134 virtual), a improvement of 15% against the native model and 8% compared to the random sampling method. Furthermore, if increasing the size of the randomly generated virtual dataset to 10000, which may lead to overly expensive labeling costs, the prediction accuracy is higher, but not noticeably higher, than our AISEL dataset with a size of 1260. As a summary, promising results in Table 3.2 suggest: (i) the proposed AISEL method efficiently incorporate physical knowledge, and therefore yields better prediction performance; (ii) with the same data size, the proposed sampling strategy, both exploring and exploring the design space, leads to a better downstream classifier than random sampling; and (iii) small AISEL dataset can achieve similar predictive performance compared to a much bigger randomly generated dataset, which reduces the labeling cost of conducting physical experiments.

3.6 Conclusion and future work

In this paper, we proposed the AISEL framework to efficiently sample a virtual dataset to incorporate complementary *physical knowledge* for small-data learning problems, with applications to manufacturing and healthcare. We first propose a novel generative invertible network (GIN), which can find the bidirectional mapping of generating virtual images and extracting the features of the actual images. We then propose a new sampling strategy, which both explores and exploits the image space to minimize the predictive uncertainty. Our AISEL method can achieve better performance in toy experiments, compared to the state-of-the-art baselines. Furthermore, in the motivating applications of aortic stenosis, our method lowers the labeling cost by 90% while achieving a 15% improvement in prediction accuracy.

Looking ahead, we are pursuing several directions for future research. From a methodological point of view, we are interested in other approaches to incorporating physical knowledge. Methods in [47, 132] appear to be attractive options. In the application point of view, further study of predicting post-surgical blood pattern is of interest. While our method can still be used, the difficulties lie in the physical labeling process. Tissue-mimicking 3D printing technology [24] and in-vitro studies [3] appear to be suitable.

CHAPTER 4
A CALIBRATION-FREE METHOD FOR BIOSENSING IN CELL MANUFACTURING

Chimeric antigen receptor T cell therapy has demonstrated innovative therapeutic effectiveness in fighting cancers; however, it is extremely expensive due to the intrinsic patient-to-patient variability in cell manufacturing. We propose in this work a novel calibration-free statistical framework to effectively recover critical quality attributes under the patient-to-patient variability. Specifically, we model this variability via a patient-specific calibration parameter, and use readings from multiple biosensors to construct a patient-invariance statistic, thereby alleviating the effect of the calibration parameter. A carefully formulated optimization problem and an algorithmic framework are presented to find the best patient-invariance statistic and the model parameters. Using the patient-invariance statistic, we can recover the critical quality attribute of interest, free from the calibration parameter. We demonstrate improvements of the proposed calibration-free method in different simulation experiments. In the cell manufacturing case study, our method not only effectively recovers viable cell concentration for monitoring, but also reveals insights for the cell manufacturing process.

4.1 Introduction

Cell therapy is one of the most promising new treatment approaches over the last decades, demonstrating great potential in treating cancers, including leukemia and lymphoma [133, 134]. Among those therapies, chimeric antigen receptor (CAR) T cell therapy [135, 136], involving the reprogramming of a patient's T cells to effectively target and attack tumor cells, has shown innovative therapeutic effects in clinical trials, leading to a recent approval (i.e., the treatment of CD19+ hematological malignancies, see [137]) by FDA as a new cancer

treatment modality. As illustrated in Figure 4.1, a typical CAR T cell therapy involves four steps – deriving cells from a patient, genetically modifying the cells, culturing the cells, and re-administering back to the patient. With increasingly mature gene modification technology, more and more researchers focus on the culturing step (i.e., the red box in Figure 4.1), where the goal is to substantially increase the cell amount from a small batch to one dose for delivery to the patient. However, a key challenge is the intrinsic patient-to-patient variability in the starting material, i.e., cells derived from different patients vary in their viabilities, acceptance rates of genetic modification, and reactions to culture media [138]. These variabilities introduce difficulties in cell culturing scale-up (i.e., cell manufacturing), and therefore, the current CAR T cell therapy is hindered by low scalability, labor-intensive processes, and extremely high cost [139]. To achieve high quality and acceptable vein-to-vein cost, we present in this work a statistical framework for online monitoring in cell manufacturing, which can alleviate the negative effect of the intrinsic patient-to-patient variability.

There are two reasons why a new statistical method is needed for monitoring critical quality attributes (exampled by the cell concentration) in cell manufacturing. Firstly, a *direct* measurement method for cell concentrations is not suitable in cell manufacturing [140]. Such a method typically requires experienced technicians to collect culture media, take microscopic images, and perform computation via an image-based software (e.g., ImageJ, see [141]). Therefore, it is labor-intensive, time-consuming, and may introduce contamination to the culture media. Furthermore, the direct measurement method is oftentimes *destructive* – the collected cells would be killed for taking microscopic images. Secondly, while there are *non-destructive* sensors available, these sensors need to be calibrated due to the unknown parameters in the sensing relationship [142]. For example, impedance sensors (adopted in this work, see Figure 4.2), which measure the dielectric relaxation of cell suspension, can be used to effectively estimate cell concentrations *after* the calibration of unknown electrical attributes, e.g., permittivity [143] and resistivity [144]. However, due

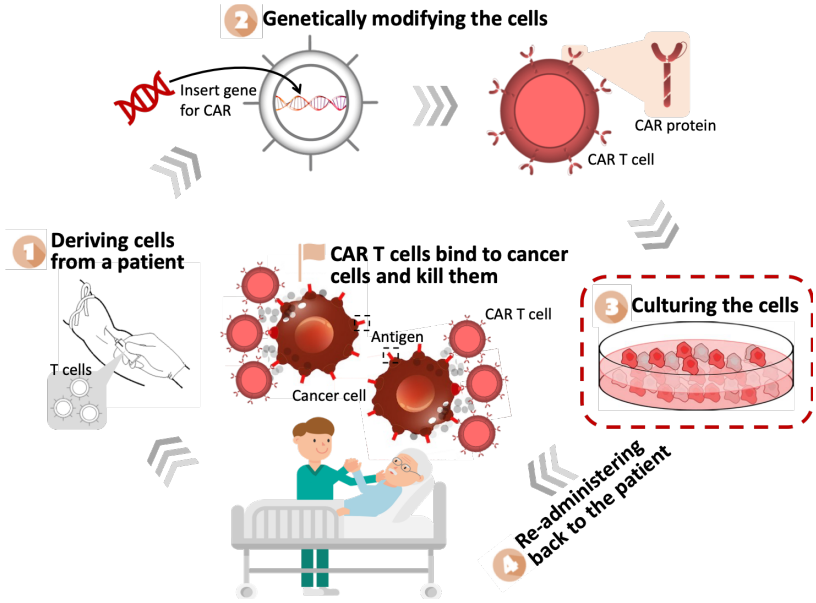

Figure 4.1: *An illustration of the four steps in a typical CAR T cell therapy. This work focuses on the cell culturing (or cell manufacturing) step, i.e., step 3.*

to the patient-to-patient variability, those electrical attributes not only are *unknown* but also *vary* among different patients, leading to difficulties in recovering cell concentrations from sensor readings.

We introduce in this work a calibration-free statistical method for online monitoring in cell manufacturing. Specifically, the intrinsic patient-to-patient variability is modeled by a patient-specific calibration parameter. We propose to use multiple sensor readings to construct a patient-*invariance* statistic, where a transformation is adopted to isolate and alleviate the effect of the calibration parameter. The constructed invariance statistic is then used to model the critical quality attribute of interest. In the *training* stage, we use the historical data to estimate a transformation and model parameters via a carefully formulated optimization problem, rather than estimate the calibration parameter as in the standard calibration problem [56, 145]. In the *monitoring* stage, we use the online sensor readings to recover the underlying critical quality attribute through the patient-invariance statistic, *free* from the calibration parameter. We demonstrate improvements of the proposed calibration-free method in both simulation experiments and a real-world case study of monitoring viable

cell concentrations in cell manufacturing. The proposed approach provides an effective way to monitor cell manufacturing, and therefore, reduces the cost for the promising CAR T cell therapy in treating cancers.

The remaining part of the article is organized as follows. In Section 4.2, we formulate the biosensing problem in cell manufacturing, with an emphasis on its challenging aspects. In Section 4.3, we present the proposed calibration-free method. A detailed simulation study and a real-world cell manufacturing case study are conducted in Sections 4.4 and 4.5, respectively. We conclude this work with future directions in Section 4.6.

4.2 Biosensing in cell manufacturing

We first describe the biosensing problem of recovering the Viable Cell Concentration (VCC) in cell manufacturing. We then discuss the key challenge – the patient-to-patient variability, and related works.

4.2.1 Impedance-based biosensing

As discussed in Section 4.1, the goal is to monitor VCC in cell manufacturing, thereby reducing the cost of the CAR T cell therapy. One state-of-the-art approach is to use biosensors to measure impedance signals, as indicators for the VCC of interest [143, 142]. As illustrated in Figure 4.2, we adopt impedance-based biosensors with a facing-electrode (FE) design [146]: Our FE biosensor consists of a pair of parallel-plate electrodes and silicone at four corners to maintain a gap between them; it would be soaked in media to monitor floating cells in between the electrodes.

With the adopted biosensors, we need a biosensing method to recover VCCs from impedance readings. From physical knowledge, we know that the impedance reading between the two electrodes reflects the cell amount due to the capacitive property of viable

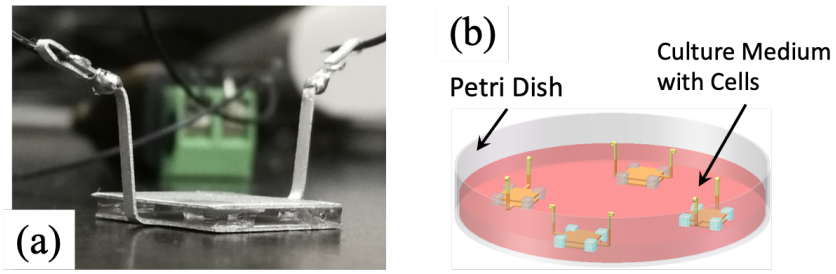

Figure 4.2: *An illustration of the adopted impedance-based biosensors in the cell manufacturing application: (a) shows a photo of the biosensor design and (b) shows the biosensing setup.*

cell membranes [147]. The sensing relationship is denoted by

$$y = f(x, \theta, \eta), \tag{4.1}$$

where y is the impedance reading, x is the underlying VCC of interest, θ denotes the sensor geometry, e.g., the gap width, and η models the underlying electrical attributes, e.g., permittivity and resistivity. Here, the relationship $f(\cdot)$ is *unknown* due to the dynamic interaction between cells and biosensors, and extremely *challenging* to simulate via computer codes considering the micro size of cells. In the proposed calibration-free method, we first learn the sensing relationship $f(\cdot)$ from the historical data of different patients (i.e., the "training" stage), and then conduct online inference of the VCC using the impedance readings for new patients (i.e., the "monitoring" stage).

4.2.2 Patient-to-patient variability

The key challenge in biosensing is that the electrical attributes η are *unknown* in both training and monitoring and *different* for each patient (also see an illustrating gravity application in Section 4.4.1). Note that it is impractical to compute η for a patient from first principle since it represents the intrinsic properties from genetic material of the patient. One popular way is to model η as a *calibration* parameter, and estimate it from the training dataset $\{y^j, x^j, \theta\}_{j=1}^J$. Existing approaches include Bayesian implementation [56], maximal likelihood estimation [148], and an interpretative l_2 optimization [145]. However, such methods are proposed

Table 4.1: *A comparison of the application scenarios of the proposed calibration-free method and other standard methods in the literature.*

Methods	$f(\cdot)$	Online η^*	Historical η^j
Inverse problem [152]	Known	Unknown	N.A.
Supervised learning [153]	Unknown	Same	Same
Calibration [56]	Unknown	Unknown	Known
Calibration-free (proposed)	Unknown	Unknown	Unknown

specifically for data fusion of computer experiments and physical experiments, where η is *available* in the former and a *constant* in the latter. Whereas in biosensing, there is *no* effective computer simulator, and electrical attributes η *vary* among physical experimental runs for different patients. In the literature, this challenge is also related to the functional calibration problem [149, 150, 151], where the calibration parameter $\eta = \eta(x)$ is modeled as a function of the input variables. In the biosensing application, however, the calibration parameter η varies among different patients yet is a constant over different input VCCs for each patient.

The biosensing problem is also related to the inverse problem in the literature [152], where one would estimate both x and η via an optimization problem. However, such a method typically assumes that the sensing relationship $f(\cdot)$ is known or can be easily learned with a *complete* data $\{y^j, x^j, \theta, \eta^j\}_{j=1}^n$, whereas, the calibration parameter η is *unknown* even in the training stage in biosensing. Furthermore, one may regard η as an additional model parameter in the unknown relationship $f(\cdot)$, and adopt a standard supervised learning scheme [153] for finding both $f(\cdot)$ and η; this implicitly assumes that η is a constant for different patients, which is *not* true in biosensing. Table 4.1 summarizes the related methods discussed; with extensive efforts in literature search, we have not found a standard method, which can be directly adopted for the biosensing problem in cell manufacturing.

4.3 Calibration-free biosensing method

We present the proposed calibration-free method in four parts. First, we discuss the sensing relationship for multiple sensors. We then introduce an invariance statistic to alleviate patient-to-patient variability. In the online monitoring stage, we use the invariance statistic to recover VCCs. In the training stage, we propose a carefully constructed optimization problem and an algorithmic framework to estimate the underlying sensing model.

4.3.1 Sensing relationship with multiple sensors

The key idea of the calibration-free method is to use *multiple* sensors to address the unknown and patient-specific calibration parameter. For the i-th sensor, we let θ_i be its geometry parameter. Consider first the sensing relationship for a given patient (or experiment). Denote $y_i[t]$ as the scalar impedance reading (more details in Section 4.5) from the i-th sensor at experimental time t. Following (4.1), we model $y_i[t] = f(x[t], \theta_i, \eta)$ via the sensing relationship $f(\cdot) : \mathbb{R}^3 \mapsto \mathbb{R}$. Here, $x[t]$ is the VCC at experimental time t, and η is the calibration parameter. It is important to note that measurements from different sensors $\{y_1[t], y_2[t], \cdots\}$ can be modeled by a *same* calibration parameter η, featuring the underlying values for electrical attributes of the specific patient's cells. This patient-specific property is the key to "canceling out" the calibration parameter using multiple sensor readings (see Section 4.3.2).

We then introduce an additional superscript j for different patients:

$$y_i^j[t] = f(x^j[t], \theta_i, \eta^j), \quad i = 1, 2, \cdots, I, \quad t = 1, \cdots T, \quad \text{and,} \quad j = 1, \cdots, J. \tag{4.2}$$

Equation (4.2) further layouts our biosensing settings with multiple sensors: (i) We assume the *homogeneity* of VCC, i.e., at given time t and a given patient j, the VCC $x^j[t]$ is the same for different senors θ_i at different locations (see Figure 4.2 (b)). This is because, in suspension cell manufacturing, the culture media is constantly stirred to ensure the

homogeneity of nutrition, and thereby VCC [154]. (ii) We use the same set of sensors with *known* parameters $\{\theta_i\}_{i=1}^{I}$ for all J patients and at different experimental time t. Those parameters are known from the fabrication process or can be easily measured from the sensors. Note that the proposed method is also effective for different sets of sensors, as long as those sensor parameters are known – the same sensor assumption is only for fabrication convenience and notation simplicity. (iii) Besides for different sensors θ_i, the calibration parameter η^j is the *same* among different measurement time t. This is because η^j models the intrinsic property of the j-th patient's cells, which typically does not change during cell manufacturing. After we clearly layout the above settings, we can then construct the patient-invariance statistic and recover the underlying VCC.

4.3.2 Invariance statistic

For notation simplicity, we drop the experimental time $[t]$ and write θ_i in the subscript in this subsection. Furthermore, we rewrite (4.2) by decomposing $f(\cdot)$ into two parts:

$$y_i^j = f_{\theta_i}(x^j, \eta^j) = \mu_i(x^j) + \delta_i(x^j, \eta^j). \tag{4.3}$$

Here, for a given sensor θ_i, $\mu_i(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ models the part of effect of VCC x on impedance reading y, *without* hampered by the patient-specific calibration parameter η; and $\delta_i(\cdot) : \mathbb{R}^2 \mapsto \mathbb{R}$ is the remaining effect of *both* x and η on y. Intuitively speaking, $\mu_i(\cdot)$ can be viewed as the *mean* process of $f(\cdot)$ by plugging in some population average of $\{\eta^j\}_{j=1}^{J}$, ignoring the patient-to-patient variability; it can also be interpreted as the physical understanding of the sensing relationship. In practice, the mean relationship $\mu_i(\cdot)$ is oftentimes known, at least to a certain degree, prior to experimentation according to the domain-specific knowledge (e.g., the known set of basis functions, see Section 4.5). On the other hand, $\delta_i(\cdot)$ is the *variability* term, i.e., how patient-to-patient variability affects the impedance reading. Such a term leads to different readings, even when the VCC x is the same. Note that in one of

the considered baseline methods (see Sections 4.4 and 4.5), we ignore $\delta_i(\cdot)$, i.e., assuming the calibration parameter is a constant; this will lead to noticeable errors when estimating x. This variability term $\delta_i(\cdot)$ is typically unknown. Such a decomposition of a mean trend and a variability term is widely assumed in different modeling methods (see, e.g., [11, 47]).

Assume for now $\delta_i(x, \eta) = \delta_i(\eta)$, which suggests that the mean relationship $\mu_i(\cdot)$ extracts all the dependency of x on y (further discussion in Section 4.3.4). In other words, $f_i(x, \eta)$ is assumed to be *separable* for each sensor θ_i:

$$y_i^j = \mu_i(x^j) + \delta_i(\eta^j). \tag{4.4}$$

Now, we construct a statistic F which is invariant to the calibration parameter η. To gain intuition, consider the following illustrating example with *known* $\delta_i(\eta) = \theta_i\eta$ for $i = 1, 2$ (see the illustration application in Section 4.4.1). If we take a log-transformation to the variability term $\log \delta_i(\eta^j) = \log \theta_i + \log \eta^j$, the effect of the calibration parameter η^j is further separated from θ_i. Therefore, by subtracting the (log-transformed) variability at different sensors, one can obtain an invariance statistic $F = \log \delta_1(\eta^j) - \log \delta_2(\eta^j) = \log \theta_1 + \log \eta^j - (\log \theta_2 + \log \eta^j) = \log(\theta_1) - \log(\theta_2)$. Note that we incorporate the patient-*specific* property of the calibration parameter when constructing the invariance statistic.

Following the above intuition yet with the *unknown* variability term $\delta_i(\cdot)$, we construct the following statistic, via a linear combination of the transformed $\delta_i(\cdot)$:

$$F(\eta^j) = \sum_{i=1}^{I} c_i \mathcal{F}[\delta_i(\eta^j)] = \sum_{i=1}^{I} c_i \mathcal{F}[y_i^j - \mu_i(x^j)], \tag{4.5}$$

for patient j with η^j. Here, $c_1, \cdots, c_I$ are pre-defined combination coefficients, and $\sum_i c_i = 0$. With a properly selected transformation $\mathcal{F}[\cdot] : \mathbb{R} \mapsto \mathbb{R}$, (4.5) gives the target *invariance* statistic $F = F(\eta^j)$. Note that here we adapt a general transformation $\mathcal{F}[\cdot]$, instead of the specific log-transformation in the above example. The transformation $\mathcal{F}[\cdot]$ would be selected so that the dependency of the invariance statistic F to η^j is minimal; a detail estimation

method for $\mathcal{F}[\cdot]$ will be discussed in Section 4.3.4.

It is important to note that we reconstruct the sensing model from (4.2) to (4.5) via the proposed invariance statistic. This is again due to the key challenge of patient-to-patient variability. Consider first using (4.2) for VCC recovery (also see the discussion in Section 4.2.2). Due to the unknown and patient-specific calibration parameter η^j, it is challenging to either learn a sensing model from training data or recover VCCs for a new patient. However, the new model (4.5) contains only the invariance statistic, and is *free* from the calibration parameter. Thanks to the properly selected transformation and the combination (see Section 4.3.4), the invariance statistic would be approximately a constant for different patients. Therefore, our new model (4.5) allows an effective estimation of the sensing relationship (only the mean part needed) similar to the standard calibration problem with a *constant* calibration parameter [145], and then a calibration-free recovery of the VCC of interest.

4.3.3 Online calibration-free recovery

We present next the method for recovering the VCC of interest x^*, in the *online monitoring* stage for a new patient denoted by $*$. At any time t, the sensor reading is denoted as $\mathcal{D}_{\text{monitor}} = \{y_i^*, \theta_i\}_{i=1}^I$ along with the unknown calibration parameter η^*. Assume for now the mean sensing relationship $\mu_i(\cdot)$ and the transformation $\mathcal{F}[\cdot]$ are known (see Section 4.3.4 for the estimation). We adopt the new sensing model (4.5) with the invariance statistic

$$F(\eta^*) = \sum_{i=1}^{I} c_i \mathcal{F}\left[y_i^* - \mu_i(x^*)\right], \tag{4.6}$$

where x^* is the target VCC. Note that the computed $F(\eta^*)$ in online monitoring is also *invariant* to the calibration parameter η^*. Therefore, the VCC of interest x^* can be recovered by minimizing the squared difference between the computed value and the underlying value

$\bar{F}$ (see Section 4.3.4 for the estimation):

$$\hat{x}^* = \underset{x^*}{\operatorname{argmin}} \left(\sum_{i=1}^{I} c_i \mathcal{F}\left[y_i^* - \mu_i(x^*) \right] - \bar{F} \right)^2 . \tag{4.7}$$

In the cell manufacturing application, we are interested in recovering a VCC curve $\hat{x}^*[t]$ over the whole manufacturing period $t = 1, \cdots, T$. To this end, we perform optimization (4.7) for T times corresponding to each experimental time t. Note that here we have not incorporated the time-dependency (or smoothness) of the recovered function $\hat{x}^*[t]$ in online recovering; one can use postprocessing methods or directly model $x^*[t]$ via a parametric form in the optimization (4.7). Readers are referred to functional data analysis literature [155, 156] for more discussion.

4.3.4 Parameter estimation

We estimate the unknown transformation $\mathcal{F}[\cdot]$, and the physical relationship $\mu_i(\cdot)$ using the training data $\mathcal{D}_{\text{train}} = \{y_i^j[t], x^j[t], \theta_i\}_{t=1}^{T}{}_{j=1}^{J}$ at hand. In our implementation, the transformation $\mathcal{F}[\cdot]$ is paramterized by the Box-Cox transformation [157, 158]

$$\mathcal{F}_\lambda[z] = \begin{cases} \log(z) & \text{if} \quad \lambda = 0 \\ \frac{(z^\lambda - 1)}{\lambda} & \text{if} \quad \lambda \neq 0 \end{cases} . \tag{4.8}$$

Note that the log-transformation in the above example is a special case in the Box-Cox transformation. Here, the Box-Cox transformation contains an unknown parameter λ. A two-parameter Box–Cox or Yeo–Johnson transformation [159] can also be used if the data are not restricted to be nonnegative. Furthermore, in this article, we will focus on the parametric transformation (4.8), but our proposed method are general and can be applied to non-parametric cases.

As for the physical relationship $\mu_i(\cdot)$, we adopt the following basis decomposition:

$$\mu_{i,\boldsymbol{\beta}}(x) = \mu_{\boldsymbol{\beta}}(x, \theta_i) = \sum_{k=1}^{K} \beta_k \phi_k(x, \theta_i). \tag{4.9}$$

Here, $\phi_k(\cdot);\ k = 1, 2, \cdots, K$ are the pre-defined basis functions and $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_K]^T$ denotes the vector of corresponding coefficients. Such a set of basis $\{\phi_k(\cdot)\}_{k=1}^{K}$ is selected by the physical knowledge of the cell manufacturing system or the observation from data.

Meanwhile, to account for the separable assumption $\delta_i(x, \eta) \approx \delta_i(\eta)$ in Section 4.3.2, we introduce slack variables $\boldsymbol{\Delta} = \{\Delta_i^j\}_{i=1}^{I}{}_{j=1}^{J}$ to account for the "goodness" of the assumption (more discussion below). Furthermore, the underlying value for the best invariance statistic $\bar{F}$ is also unknown and need to be estimated from data (see Section 4.3.3).

We propose to estimate the unknown parameters $\{\lambda, \boldsymbol{\beta}, \bar{F}, \boldsymbol{\Delta}\}$ via the following optimization problem with two penalization terms:

$$\begin{aligned}\min_{\lambda,\boldsymbol{\beta},\bar{F},\boldsymbol{\Delta}} l_\alpha(\lambda, \boldsymbol{\beta}, \bar{F}, \boldsymbol{\Delta}) = \min_{\lambda,\boldsymbol{\beta},\bar{F},\boldsymbol{\Delta}} \sum_{t,j} \left[\sum_{i=1}^{I} c_i \mathcal{F}_\lambda[y_i^j[t] - \mu_{\boldsymbol{\beta}}(x^j[t], \theta_i)] - \bar{F} \right]^2 \\ + \alpha_1 \sum_{i,j,t} \left[|y_i^j[t] - \mu_{\boldsymbol{\beta}}(x^j[t], \theta_i)|_1 - \Delta_i^j \right]^2 + \alpha_2 |\boldsymbol{\beta}|_1.\end{aligned} \tag{4.10}$$

Here, α_1 and α_2 are two penalization coefficients, and $|\cdot|_1$ denotes the vector l_1 norm.

The main objective term (i.e., the first term) in (4.10) is for achieving the best patient-invariance property of the constructed statistic F. Specifically, we minimize the mean-squared error (MSE) of its underlying truth $\bar{F}$ and the computed value from data. This is equivalent to modeling the patient-invariance statistic F^j for each patient as i.i.d. random draws from a normal distribution $\mathcal{N}(\bar{F}, \sigma^2)$ with mean $\bar{F}$ and variance σ^2. Moreover, the first penalization term in (4.10) is for the separable assumption $\delta_i(x, \eta) \approx \delta_i(\eta)$. Here, we minimize the corresponding MSE of the set $\{\delta_i(x^j[t], \eta^j)\}_{t=1}^{T}$ to the underlying truth Δ_i^j, for each sensor i and patient j. Similarly, this can also be viewed as modeling $\delta_i(\cdot, \eta^j)$ via

Algorithm 5 The BCD algorithm for parameter estimation (4.10)

- Set initial values $\lambda \leftarrow 0$ and $\boldsymbol{\beta} \leftarrow \mathbf{1}_K$
- Set $I \times J \times T \times K$ tensor $\boldsymbol{\Phi}$ with each element $\Phi_{ijtk} \leftarrow \phi_k(x^j[t], \theta_i)$

repeat

Optimizing F and $\boldsymbol{\Delta}$:

- Set $I \times J \times T$ tensor $\mathbf{D}$ with each element $D_{ijt} \leftarrow y_i^j[t] - \sum_k \Phi_{ijtk}\beta_k$
- Update $\bar{F} \leftarrow \sum_{ijt} c_i \mathcal{F}_\lambda[D_{ijt}]/J/T$
- Update $\Delta_i^j \leftarrow \sum_t |c_i \mathcal{F}_\lambda[D_{ijt}|_1/T$ for $i = 1, \cdots I$ and $j = 1, \cdots J$

Optimizing λ:

- Set $l_0(\lambda, \boldsymbol{\beta}, \bar{F}, \boldsymbol{\Delta}) \leftarrow l_\alpha(\lambda, \boldsymbol{\beta}, \bar{F}, \boldsymbol{\Delta})$ with $\alpha_1 = \alpha_2 = 0$
- Update $\lambda \leftarrow \text{argmin}_\lambda \, l_\alpha(\lambda, \boldsymbol{\beta}, \bar{F}, \boldsymbol{\Delta})$

Optimizing $\boldsymbol{\beta}$:

- Update $\boldsymbol{\beta} \leftarrow \text{argmin}_{\boldsymbol{\beta}} \, l_\alpha(\lambda, \boldsymbol{\beta}, \bar{F}, \boldsymbol{\Delta})$ with the L-BFGS method

until λ, $\bar{F}$, $\boldsymbol{\Delta}$ and $\boldsymbol{\beta}$ converge

- **return** λ, $\bar{F}$, $\boldsymbol{\Delta}$ and $\boldsymbol{\beta}$

i.i.d. normal random variables; the corresponding penalization α_1 can then be interpreted as the ratio between the variances of the two normal distributions. Finally, the second penalization term $\alpha_2|\beta|_1$ is for basis selection, similar to the widely used LASSO method in the literature [160, 161]. This is because, in the cell manufacturing application, one would only have an intuitive understanding of the sensing relationship; we will collect a set of basis functions from experience and select the suitable ones via this penalization.

From the duality of the optimization problem, (4.10) can also be viewed as unpenalized log-likelihood combining both normal random variables with a sparsity constraint $\|\boldsymbol{\beta}\|_1 \leq \gamma$ [162]. The parameter α_1 sets the variance ratio between the two random variables, and α_2 controls the desired sparsity level as parameter γ. Since the objective is to obtain a high recovery accuracy of the VCC of interest, α_1 and α_2 would be tuned using cross-validation techniques [36].

Consider now estimating the parameters $\{\lambda, \boldsymbol{\beta}, \bar{F}, \boldsymbol{\Delta}\}$ via optimization (4.10) for fixed $\alpha_1 > 0$ and $\alpha_2 > 0$. We propose to use the following Blockwise Coordinate Descent (BCD) optimization algorithm, described below. First, assign initial values for $\{\lambda, \boldsymbol{\beta}, \bar{F}, \boldsymbol{\Delta}\}$. Next, iterate the following three steps until the convergence is achieved: (i) for fixed λ and $\boldsymbol{\beta}$,

update $\{\bar{F}, \boldsymbol{\Delta}\}$ via the following estimates from first-order conditions

$$\hat{\bar{F}} = \frac{1}{JT} \sum_{i,j,t} c_i \mathcal{F}_\lambda \left[y_i^j[t] - \mu_\beta(x^j[t], \theta_i) \right], \quad \hat{\Delta}_i^j = \frac{1}{T} \sum_{t=1}^{T} |y_i^j[t] - \mu_\beta(x^j[t], \theta_i)|_1; \quad (4.11)$$

(ii) for fixed $\boldsymbol{\beta}$ and $\bar{F}$, update the transformation parameter λ ignoring the two penalization terms; and (iii) for fixed λ, $\bar{F}$ and $\boldsymbol{\Delta}$, optimize for $\boldsymbol{\beta}$ via numerical line search methods, e.g., L-BFGS algorithm [37]. The full optimization procedure is provided in Algorithm 5. Since (4.10) is a non-convex optimization problem, the proposed BCD algorithm only converges to a stationary solution [38]. Because of this, we suggest performing multiple runs of Algorithm 5 with random initializations for each run, then taking the converged estimates for the run with smallest objective function. These runs should be performed in parallel if possible, to take advantage of the parallel computing capabilities in many computing systems.

It is important to note the difference between the *training* stage in our calibration-free method and the *calibration* stage in the standard calibration problem [56]. In calibration methods, the calibration parameter η, assumed to be a constant, is directly estimated from the training set. This can be viewed as estimating a population average of the historical $\{\eta^j\}_{j=1}^J$, which would *not* be helpful in our cell manufacturing application. Due to the patient-to-patient variability, the calibration parameter η^* corresponding to the new patient can be completely different from the historical average value. In contrast, our calibration-free method adopts a patient-invariance statistic F, constructed from multiple sensor readings, to alleviate this patient-to-patient variability. In our training setup, we learn the unknown mean relationship $\mu_i(\cdot)$ and the transformation $\mathcal{F}[\cdot]$, which provide the best patient-invariance statistic F. We can then use the invariance statistic F to effectively recover VCCs via (4.7), free from the patient-specific calibration parameter η^*.

4.4 Simulation study

A detailed simulation study is conducted in this section. We first look into a toy application of recovering gravitational acceleration coefficients, to show the applicability of the proposed method. We then discuss more simulation experiments with different sensing relationships.

4.4.1 A gravity application

Consider the following toy application, where the goal is to recover the gravitational acceleration coefficient x, for a different planet. As illustrated in Figure 4.3, we drop a ball and measure the traveling distance y of the ball after a certain period of time θ. From the physical knowledge, we have the relationship

$$y = f(x, \theta, \eta) = \frac{1}{2}x\theta^2 + \eta\theta. \tag{4.12}$$

Here, y is the traveling distance measured by, e.g., taking a photo, η is the initial velocity of the ball, and θ is the time period between dropping the ball and taking the photo. Suppose the ball is dropped by an engineer, meaning that the initial velocity η is non-zero and changes among different drops. With the collected data $\{y, \theta\}$, typically, one cannot recover the gravitational acceleration x even with the known relationship (4.12). This is because the initial velocity η is also *unknown*. The key idea of the proposed calibration-free method is to take *multiple* photos at different time $\theta_i; i = 1, \cdots, I$. Therefore, more data $\{y_i, \theta_i\}_{i=1}^I$ is collected with the *same* initial velocity η. We can then use the proposed invariance statistic and Algorithm 5 to "cancel out" η and conduct inference on the gravitational acceleration x of interest.

The setup for recovering the gravity coefficient is as follows. We set the number of photos $I = 3$, with parameters (i.e., the time of taking photos) $\{\theta_i\}_{i=1}^3 = \{0.5, 1, 1.5\}$. A historical dataset $\{y_i^j, x^j; \eta^j\}_{j=1}^J$ of size $J = 100$ is generated with calibration parameters (i.e., initial speed, *unknown*) $\eta^j \sim Unif(1, 3)$, gravity coefficients $x^j \sim \mathcal{N}(9.8, 1^2)$, and

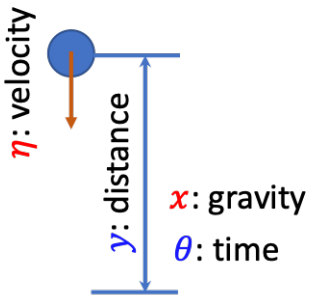

Figure 4.3: *An illustration and notations of the toy application of recovering the gravitational acceleration coefficient.*

each sensor reading (i.e., traveling distance) y_i^j simulated by (4.12) with an additional i.i.d. measurement error following $\mathcal{N}(0, 0.4^2)$. To test the recovery accuracy, we let the underlying truth $x^* = 9.8$, $\eta^* \sim Unif(1, 3)$ randomly generated, and y^* obtained by (4.12) with the same measurement error.

The proposed calibration-free method (via Algorithm 5) is applied to find the best transformation $\mathcal{F}_{\hat{\lambda}}[\cdot]$ and then recover the gravitational acceleration $\hat{x}^*$. The linear combination coefficients are $\{c_i\}_{i=1}^3 = \{1, 1, -2\}$, and the set of candidate basis functions is $\Phi = \{x, \theta, \theta x, \theta x^2, \theta^2 x\}$. The proposed calibration-free method is repeated, with newly generated test data $\{y^*, x^* = 9.8; \eta^*\}$, for 20 times.

We consider the following two baseline methods (also see Table 4.1). First, we implement the supervised learning setting [153], i.e., assuming the calibration parameter $\eta = \eta^j = \eta^*$ is the *same* in both training stage and monitoring stage. Such an assumption is *not* true in the considered cell manufacturing application. Specifically, we use the historical dataset $\{y_i^j, x^j\}_{j=1}^{100}$ to estimate an $\bar{\eta}$ and a relationship $g(x, \theta) := f(x, \theta; \bar{\eta})$, which would then be used to recover $\hat{x}^*$. Here, we use the same set of basis functions Φ for $g(\cdot)$, and a similar optimization scheme with LASSO type penalization [160] for estimating the coefficients $\boldsymbol{\beta}$. This method is referred to as "SameCal". The other baseline method is the standard l_2 calibration method suggested by [145]. In order to adopt this method, we need to assume that the calibration parameters $\{\eta^j\}_{j=1}^{100}$ in the historical data are *known*, which is *not* true in reality. Therefore, we refer this method as "Oracle". To estimate the sensing

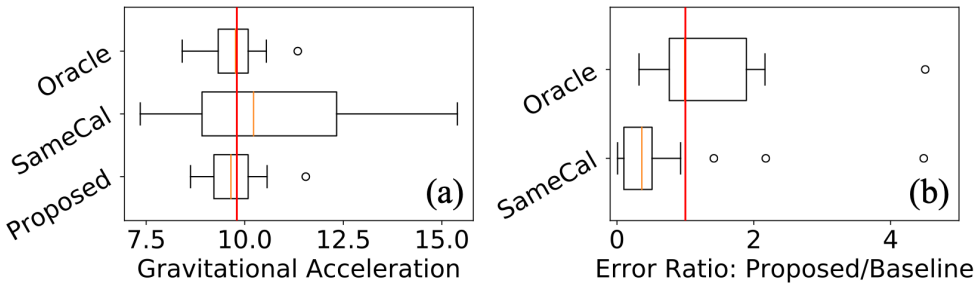

Figure 4.4: *Results of the gravity application: (a) shows the recovered gravitational acceleration by the three considered methods. The red line marks the underlying truth $x^* = 9.8$. (b) shows the boxplots of absolute error ratios between the proposed method and baseline baseline methods. The red line marks the ratio of 1.0.*

relationship $f(\cdot)$, we adopt the set of basis functions $\Phi_o = \{x, \theta, \theta x, \theta x^2, \theta^2 x, \eta x, \eta\theta, \eta x\theta\}$ and a similar optimization scheme with LASSO type penalization. Both baseline methods are implemented to recover x^* of interest via minimizing the squared difference similar to (4.7), using the same 20 simulated test data.

Figure 4.4 (a) shows the boxplots of the estimated $\hat{x}^*$ using the proposed calibration-free method and the two baseline methods. The red line indicates the ground truth value $x^* = 9.8$. Among the three methods, the Oracle baseline preforms the best since it queries *additional* information of the calibration parameter in the training stage, which is again *not* feasible in reality. We notice that the proposed calibration-free method performs almost as good as Oracle. It can accurately recover the true value, with the mean over the 20 estimates 9.7 and a relatively small variance, whereas for the SameCal baseline, the mean for 20 estimates is 10.8 and a noticeable bigger variance is observed.

We also conduct a pairwise comparison over the 20 test repetitions. Figure 4.4 (b) shows the boxplots of absolute error ratios of the proposed method over the two baseline methods, with the red line indicating the ratio of 1.0. We notice that the proposed method is only slightly worse than Oracle; this is impressive since our method do *not* query the underlying calibration parameter in the training stage. Furthermore, the proposed calibration-free method is noticeably better in recovering the true x^* compared to SameCal. More specifically, our method outperforms SameCal with smaller errors in 17 estimates over

20 test runs in total. This is not surprising since the calibration-free method, utilizing the patient-invariance statistic, can address the patient-specific calibration parameters η^*.

4.4.2 More experiments

Here, we conduct more experiments on the proposed calibration-free method. Specifically, we consider the following four underlying sensing relationships $f_k(x, \theta, \eta)$:

1. $f_1(x, \theta, \eta) = x\theta + \eta\theta^2$;
2. $f_2(x, \theta, \eta) = 3x + 2\theta x + x\theta\eta$;
3. $f_3(x, \theta, \eta) = \theta x + x^2 + \theta\eta^2 + \sqrt{\theta}\eta^2 + \sqrt{x}\eta/4$;
4. $f_4(x, \theta, \eta) = \sin(x) + (x + \eta)^\theta + \frac{x}{\theta+\eta}$.

Note that function $f_1(\cdot)$ is quite similar to the sensing relationship (4.12) in the gravity application in Section 4.4.1. For functions $f_2(\cdot)$ and $f_3(\cdot)$, we notice the existence of interaction terms between x and η, which means the separable assumption $\delta_i(x, \eta) = \delta_i(\eta)$ in Section 4.3.2 does not hold. However, $f_2(\cdot)$ and $f_3(\cdot)$ can still be approximately represented by the adopted set of basis functions Φ. For function $f_4(\cdot)$, it is quite complex, and cannot be represented by Φ. We test all four functions, using the proposed calibration-free method and the same two baseline methods – SameCal and Oracle – introduced in Section 4.4.1. The detailed test procedure is the same as that in Section 4.4.1.

Figure 4.5 shows the boxplots of the absolute error ratios of the proposed method over the baseline SameCal method (a) and the baseline Oracle method (b), under all four underlying sensing functions $f_k(x, \theta, \eta);\ k = 1, \cdots, 4$. We notice the error ratios in Figure 4.5 (a) are mostly smaller than 1.0, indicting that the proposed method can achieve smaller errors compared to SameCal. This is because the assumption of constant calibration parameter in SameCal does *not* hold in cell manufacturing application (and thereby in this simulation study), whereas, our calibration-free method can address the patient-specific calibration

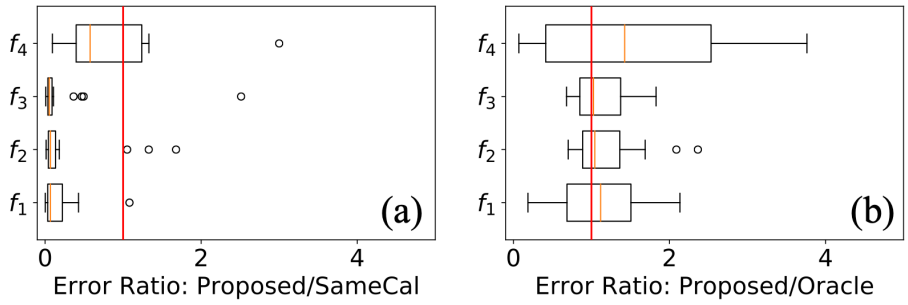

Figure 4.5: *Boxplots of error ratios between the proposed method and the considered baselines, under different sensing relationships. The red line marks the ratio of 1.0, indicating the baseline method achieves the same accuracy as the proposed method.*

parameter via a proper combination of multiple sensor readings. Moreover, compared to Oracle, the proposed method is only slightly worse. This shows the good performance of our calibration-free method: Though we do *not* know the values of the calibration parameter, we *can* recover the underlying parameter of interest similar to the Oracle baseline, where whose values *are* assumed accessible. Finally, we notice that for the sensing relationship $f_4(\cdot)$, while the proposed method adopts an inappropriate basis decomposition Φ, it still outperforms SameCal. This is again because the proposed calibration-free method introduces the invariance statistic to alleviate the effect of patient-to-patient variability, and therefore, shows improved performances in recovering the quantity of interest.

4.5 Cell manufacturing case study

In this section, we apply the proposed calibration-free method to the motivating case study of cell manufacturing. As discussed in Sections 4.1 and 4.2, we are interested in recovering and monitoring the Viable Cell Concentration (VCC) $x[t]$ at different experimental time t. This is because the goal of cell manufacturing is to culture a small batch of cells to a significant amount, for an effective re-administering in the CAR T cell therapy (see Figure 4.1).

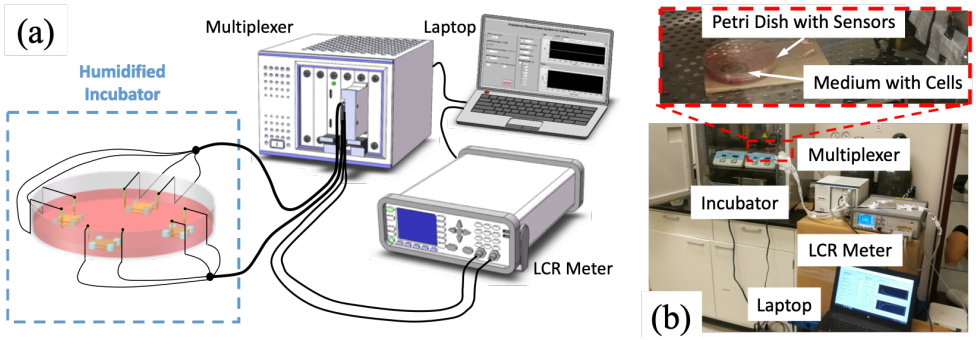

Figure 4.6: *The cell manufacturing application: (a) an illustration and (b) the actual experimental setup with an emphasis on the impedance measurement part.*

4.5.1 Experimental setup

In our experiment, human leukemic T cells (Jurkat E6-1; American Type Culture Collection, ATCC®) are cultured in ATCC-formulated culture medium (RPMI-1640; GE Healthcare) with 10% fetal bovine serum, 2 mM L-glutamine, 10 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 1 mM sodium pyruvate, 4500 mg/L glucose, and 1500 mg/L sodium bicarbonate in a 75 cm^2 petri dish (Nunc™ EasYFlask™; ThermoFisher Scientific™). The cells are cultured in a humidified incubator controlled at 37°C and 5% CO^2, and the culture media is pre-heated to avoid the temperature effect on the impedance measurement [163].

The impedance measurements are obtained by our electric cell-substrate impedance sensing. Figure 4.6 illustrates the experimental setting for the impedance measurement. Here, we use $I = 4$ sandwich shape 3D impedance sensors, consisting a pair of parallel-plate electrodes and PDMS (Sylgard 184, Tow Corning) to maintain a gap between two electrodes (see Figure 4.2 (b)). In our experimental setup, the geometry parameter θ of the sensors is the edge length of the electrode pads, and $\{\theta_i\}_{i=1}^4 = \{8mm, 10mm, 12mm, 14mm\}$. Impedance measurements are conducted by an LCR meter (E4980AL; Keysight Technologies) with a sinusoidal signal of 22 mVrms under multiple frequencies ranging from 500 Hz to 100 kHz. We let impedance measurement y be the relaxation strength computed from the raw impedance readings over frequencies, i.e., the difference between permittivity values of

Table 4.2: *Cross-validation errors of the recovered VCCs for the cell manufacturing case study, using the proposed calibration-free method and the baseline SameCal method.*

	Exp 1	**Exp 2**	**Exp 3**	**Exp 4**	**Exp 5**	**Mean**
Proposed	**0.092**	0.270	**0.080**	**0.379**	**0.590**	**0.282**
SameCal	0.500	**0.174**	0.328	0.742	0.760	0.501

high-frequency end and low-frequency end of the dielectric relaxation process, for its high dependency to the VCC of interest [147]. The measurement is taken every 15 minutes for around 30-35 hours. This is because typically after 35 hours, we have to change the culture media and expand the cells to a bigger cell culture flask, which would inevitably interrupt the online monitoring. This results in an online monitoring dataset $\mathcal{D}_{\text{monitor}} = \{y_i^j[t], \theta\}_{t=1}^{T}{}_{j=1}^{J}$ with the underlying VCCs to be recovered.

The ground truth VCCs are obtained by an automated cell counter (TC20™; Bio-Rad Laboratories, Inc.), and the concentration is maintained between 1×10^5 and 1×10^6 cells/mL. Multiple repetitions are performed, with the averaged value reported as the underlying VCC x. Note that the measurement procedure is not only labor-intensive but may also introduce contamination to the culture media (see Section 4.1). We will only measure VCC around six times per cell culture experiment, which leads to a much smaller ($\tau \ll T$) yet full training dataset $\mathcal{D}_{\text{train}} = \{y_i^j[t], x^j[t], \theta\}_{t=1}^{\tau}{}_{j=1}^{J}$.

We conduct the cell culturing for $J = 5$ experiments, each with different starting materials, i.e., different calibration parameters η^j. We use the same $I = 4$ sensors $\{\theta_i\}_{i=1}^4 = \{8mm, 10mm, 12mm, 14mm\}$ for the $J = 5$ experiments. In experiment j, we measure and compute the relaxation strength of impedance $y_i^j[t]$ for each sensor i, at different time t (every 15 minutes, $T \approx 130$ in total). Meanwhile, we measure the ground truth VCC $x^j[t]$ with a much lower resolution ($\tau \approx 6$ in total).

4.5.2 Cross validation of viable cell concentration

For the collected training dataset $\mathcal{D}_{\text{train}}$, we first preform a cross-validation test [36] on the recovered VCCs. More specifically, we apply the proposed calibration-free method (via

Algorithm 5) to learn the sensing relationship using four out of five experiments, and then recover VCC $\hat{x}^j[t]$ for the remaining experiment via (4.7). We let the linear combination coefficients $\{c_i\}_{i=1}^4 = \{1, -1, 1, -1\}$ and select the same set of basis functions Φ as in Section 4.4.1. Furthermore, a log-transformation on VCCs is performed prior to the analysis. We consider SameCal (see Section 4.4) as the baseline method. Such a method introduces an additional assumption that the calibration parameter is a constant. Note that the Oracle baseline cannot be adopted here since the actual values of the calibration parameter are always *unknown*, which is the key motivation of the proposed calibration-free method (also see Section 4.2.2 and Table 4.1).

Table 4.2 shows the absolute errors of the recovered log VCCs when the ground truth VCCs are measured in each experiment. We observe that the proposed method outperforms the baseline SameCal method four experiments out of five. Furthermore, the mean error of the five experiments by the proposed method is 0.282, which is almost two times smaller than that of 0.501 by SameCal. This is due to the fact that the calibration parameter, which models the patient-to-patient variability, is *not* a constant in cell manufacturing [138]; the proposed calibration-free method properly addresses this variability via the construction of the patient-invariance statistic.

4.5.3 Online recovery of viable cell concentration

We then perform VCC recovery on the online monitoring set $\mathcal{D}_{\text{monitor}}$. Here, the sensing relationship is estimated using all five experiments in the training set $\mathcal{D}_{\text{train}}$. Figure 4.7 shows the two recovered log VCC curves over the whole culture time $\log \hat{x}^j[t]$, via the proposed calibration-free method (in red line) and the baseline SameCal method (in green dash line). The ground truth (log) VCC measurements in $\mathcal{D}_{\text{train}}$ are also plotted in black dots. We see that the proposed method recovers a meaningful estimation of VCC. The recovered $\log \hat{x}^j[t]$ increases approximately linear over the culture time t, indicating $\hat{x}^j[t]$ growths exponentially in time; this matches the preliminary understanding in the cell

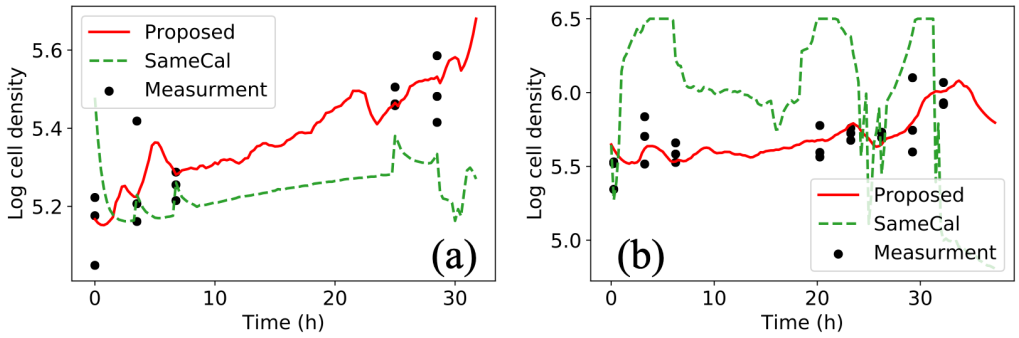

Figure 4.7: *The recovered VCC over time of two cell manufacturing experiments under the two considered methods. The ground truth VCC measurements are shown in dots.*

culture literature [154]. Furthermore, the recovered $\hat{x}^j[t]$ approximately passes through the ground truth measurements. However, due to the huge patient-to-patient variability, the baseline SameCal method struggles in either passing through the ground truth experiments or providing reasonable estimates of VCC curves. Our calibration-free method, adopting the patient-invariance statistic, appears to alleviate such variability well.

The proposed calibration-free method can also provide important biological insights for cell growth in cell manufacturing. From Figure 4.7 (b), we notice a decrease in the VCC curve at around hour 32. This may be due to the lack of nutrition in the media since the culture media typically needs to be changed after 30 hours. Furthermore, we observe from Figure 4.7 that the VCC curves decrease slightly in the first two hours in cell manufacturing. This may be because of the lack of viability of the cells at the beginning of the culture process – though we have already thaw cells and stood them still for several minutes, it seems that a certain portion of cells still do not gain full viability and die soon. As a result, we suggest standing the cell still longer for future experiments. Last but not least, we notice a small VCC decrease when conducting the ground truth VCC measurements. One reason for this is that the measurement itself is not in-line and needs to contact the culture media; it may introduce contamination, and therefore, kill a small portion of cells [154]. In contrast, the proposed calibration-free biosensing method, together with impedance-based biosensors, provides an in-line, non-destructive, and non-contact way for VCC monitoring

in cell manufacturing.

4.6 Conclusion

In this work, we propose a new calibration-free method for monitoring viable cell concentration in cell manufacturing, which is a critical component in the promising CAR T cell therapy. The key challenge here is the patient-to-patient variability in the initial culturing material, leading to poor performances in recovering viable cell concentrations via existing methods. We propose to use multiple impedance-based biosensors with different geometries and an associated calibration-free statistical framework for online recovery of viable cell concentrations. Specifically, we model the patient-to-patient variability via a patient-specific calibration parameter. We then construct a patient-invariance statistic, which uses a transformation and a linear combination of sensor readings to alleviate the effect of the calibration parameter. In the training stage, we learn the best transformation and the sensing relationship via a carefully formulated optimization problem. In the online monitoring stage, viable cell concentrations can be recovered via the invariance statistic, free from the patient-specific calibration parameter. We then apply the proposed calibration-free method in different simulation experiments and a real-world case study of cell manufacturing, where the proposed method demonstrates substantial improvements against the existing methods. Therefore, we believe the proposed calibration-free method can play an essential role in cell manufacturing and reduce the cost of the promising CAR T cell therapy.

Looking ahead, there are several interesting directions for future exploration. To begin with, a more thorough analysis of impedance-based sensors can be conducted, with a detailed comparison of sensitivity using different experimental settings such as sensor geometries and electrode materials. Moreover, we adopt in this work a parametric sensing relationship and a heuristic approach for parameter estimation. This is mainly due to the already improved performance compared to the baseline methods. A more flexible, and non-parametric Gaussian process regression method [6, 164] with a rigorous likelihood-

based parameter estimation scheme may lead to further improvements in recovering viable cell concentrations, as well as other critical quality attributes. Finally, micro cameras can also be used in cell manufacturing. Therefore, we are also interested in monitoring cell manufacturing based on cell morphology. In this case, physics-informed deep learning frameworks in the literature [165, 166, 42] appear to be suitable for recovering critical quality attributes in cell manufacturing.

Appendices

APPENDIX A

APPENDIX FOR CHAPTER 1

A.1 Proof of Theorem 1

For any $n \in \mathbb{N}, c_1, \cdots, c_n \in \mathbb{R}$ and function $I_1(\cdot), \cdots, I_n(\cdot) \in \mathcal{I}$, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \rho(I_i(\cdot), I_j(\cdot)) \tag{A.1}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \exp\left(-\int \theta(\xi)\left(\left|\hat{I}_i(\xi)\right| - \left|\hat{I}_j(\xi)\right|\right)^2 d\xi\right). \tag{A.2}$$

Note that the Fourier transform $\mathcal{F} : \mathcal{I} \mapsto \mathcal{I}$, where $\mathcal{I}$ is the space of integrable functions $I(\cdot) : \mathbb{R} \mapsto \mathbb{C}$, has a unique inverse $\mathcal{F}^{-1} : \mathcal{I} \mapsto \mathcal{I}$. Denote the standard Gaussian kernel as $K(\cdot,\cdot) : |\mathcal{I}| \times |\mathcal{I}| \mapsto \mathbb{R}$,

$$K\left(F_1(\cdot), F_2(\cdot)\right) = \exp\left(-\int \theta(t)(F_1(t) - F_2(t))^2 dt\right), \tag{A.3}$$

where $|\mathcal{I}|$ is the space of integrable functions $F(\cdot) : \mathbb{R} \mapsto \mathbb{R}$. Since $K(\cdot,\cdot)$ is a positive definite kernel, for the selected n and $c_1, \cdots, c_n$ in Equation (A.1), and any function $F_1(\cdot), \cdots, F_n(\cdot) \in \mathcal{I}$, we have,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(F_i(\cdot), F_j(\cdot)) \geq 0. \tag{A.4}$$

Now let $F_k(\cdot) = \left|\hat{I}_k(\cdot)\right|$, where $k = 1, 2, \cdots, n$. This is possible because $\hat{I}_k(\cdot) \in \mathcal{I}$ and

therefore $\left|\hat{I}_k(\cdot)\right| \in \mathcal{I}$. Thus, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K\left(\left|\hat{I}_i(\cdot)\right|, \left|\hat{I}_j(\cdot)\right|\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \exp\left(-\int \theta(\xi)\left(\left|\hat{I}_i(\xi)\right| - \left|\hat{I}_j(\xi)\right|\right)^2 d\xi\right) \geq 0 \tag{A.5}$$

In other words,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \rho(I_i(\cdot), I_j(\cdot)) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(F_i(\cdot), F_j(\cdot)) \geq 0, \tag{A.6}$$

i.e., the proposed SpeD correlation function is positive semi-definite.

APPENDIX B

APPENDIX FOR CHAPTER 2

B.1 Single-fidelity ICMSE design criterion

B.1.1 A useful intermediate derivation

We present first a simplified expression for the design criterion (2.7), which will aid in later derivations.

Let Y'_{n+1} be the latent response at $\mathbf{x}_{n+1}$ *prior* to censoring, and $Y_{n+1} = Y'_{n+1}(1 - \mathcal{I}(\mathbf{x}_{n+1})) + c\ \mathcal{I}(\mathbf{x}_{n+1})$ be the corresponding response *after* censoring, with c the right-censoring limit. Here, we define the censoring indicator function:

$$\mathcal{I}(\mathbf{x}_{n+1}) = \mathbb{1}_{\{Y'_{n+1} \geq c\}} = \begin{cases} 0 & \text{if} \quad Y'_{n+1} \geq c \\ 1 & \text{if} \quad Y'_{n+1} < c \end{cases}.$$

We can now define the probability of censoring at potential input point $\mathbf{x}_{n+1}$ as $\lambda = \lambda(\mathbf{x}_{n+1}) = \mathbb{P}[\mathcal{I}(\mathbf{x}_{n+1}) = 1|\mathcal{Y}_n] = \mathbb{P}[Y'_{n+1} \geq c|\mathcal{Y}_n]$.

Let $\text{CMSE}(\mathbf{x}_{n+1}, \mathbf{x}_{\text{new}})$ be the integrand of the proposed criterion (2.7). One can decompose this integrand via the total variance formula:

$$\text{CMSE}(\mathbf{x}_{n+1}, \mathbf{x}_{\text{new}}) = \text{Var}[\xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n] - \text{Var}_{Y_{n+1}|\mathcal{Y}_n}\left[\mathbb{E}\left(\xi(\mathbf{x}_{\text{new}})|Y_{n+1}, \mathcal{Y}_n\right)\right].$$

Denote the first term $\text{Var}[\xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n] = \sigma^2_{\text{new}}$. As for the second term, we compute variance of the random variable $Z = \mathbb{E}\left(\xi(\mathbf{x}_{\text{new}})|Y_{n+1}, \mathcal{Y}_n\right)$ by conditioning on the censoring indicator function $\mathcal{I} = \mathcal{I}(\mathbf{x}_{n+1})$:

$$\text{Var}[Z] = \mathbb{E}_{\mathcal{I}}[\text{Var}(Z|\mathcal{I})] + \text{Var}_{\mathcal{I}}[\mathbb{E}(Z|\mathcal{I})].$$

Consider first the expected variance term $\mathbb{E}_{\mathcal{I}}[\text{Var}(Z|\mathcal{I})]$. Since Z is a constant when censoring occurs (i.e., $\mathcal{I} = 1$), the first term becomes:

$$\begin{aligned}\mathbb{E}_{\mathcal{I}}[\text{Var}(Z|\mathcal{I})] &= (1-\lambda)\text{Var}_{Y_{n+1}|\mathcal{I}=0}[\mathbb{E}(\xi(\mathbf{x}_{\text{new}})|Y_{n+1})] + \lambda \times 0 \\ &= (1-\lambda)\left(\text{Var}[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} \geq c] - \mathbb{E}_{Y_{n+1}|Y'_{n+1}<c}[\text{Var}(\xi(\mathbf{x}_{\text{new}})|Y_{n+1})]\right),\end{aligned} \tag{B.1}$$

where the condition on data $\mathcal{Y}_n$ is omitted for brevity. Consider next the variance of expectation term $\text{Var}_{\mathcal{I}}[\mathbb{E}(Z|\mathcal{I})]$. Note that the random variable $\mathbb{E}[Z|\mathcal{I}]$ follows a two point distribution. Hence, the second term becomes:

$$\begin{aligned}\text{Var}_{\mathcal{I}}[\mathbb{E}(Z|\mathcal{I})] &= \lambda(1-\lambda)\left(\mathbb{E}\left[\xi(\mathbf{x}_{\text{new}})|\mathcal{I}=1\right] - \mathbb{E}_{Y_{n+1}|\mathcal{I}=0}\left[\mathbb{E}(\xi(\mathbf{x}_{\text{new}})|Y_{n+1})\right]\right)^2 \\ &= \lambda(1-\lambda)\left(\mathbb{E}\left[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} \geq c\right] - \mathbb{E}\left[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} < c\right]\right)^2,\end{aligned} \tag{B.2}$$

where the condition on data $\mathcal{Y}_n$ is again omitted for brevity. Putting these together, we have the following expression for $\text{CMSE} = \text{CMSE}(\mathbf{x}_{n+1}, \mathbf{x}_{\text{new}})$:

$$\begin{aligned}\text{CMSE} = \sigma^2_{\text{new}} &- \lambda(1-\lambda)\left(\mathbb{E}\left[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} \geq c\right] - \mathbb{E}\left[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} < c\right]\right)^2 \\ &- (1-\lambda)\left(\text{Var}[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} \geq c] - \mathbb{E}_{Y_{n+1}|Y'_{n+1}<c}[\text{Var}(\xi(\mathbf{x}_{\text{new}})|Y_{n+1})]\right).\end{aligned} \tag{B.3}$$

which again is the integrand of the proposed criterion $\text{ICMSE}(\mathbf{x}_{n+1})$.

B.1.2 Proof of Theorem 2

Suppose no censoring in training data, i.e., $\mathcal{Y}_n = \{\mathbf{y}_o\}$. Using the conditional mean and variance expressions (5) and (6) for standard GP regression, we have:

$$\begin{bmatrix} Y'_{n+1} \\ \xi(\mathbf{x}_{\text{new}}) \end{bmatrix} \Big| \mathcal{Y}_n \sim \mathcal{N}\left(\begin{bmatrix} \mu_{n+1} \\ \mu_{\text{new}} \end{bmatrix}, \begin{bmatrix} \sigma^2_{n+1} & \rho\sigma_{n+1}\sigma_{\text{new}} \\ \rho\sigma_{n+1}\sigma_{\text{new}} & \sigma^2_{\text{new}} \end{bmatrix}\right).$$

Here, the predictive means are:

$$\mu_{n+1} = \mathbb{E}[Y'_{n+1}|\mathcal{Y}_n] = \mu_\xi + \boldsymbol{\gamma}_{n,n+1}^T \boldsymbol{\Gamma}_n^{-1}(\mathbf{y}_o - \mu_\xi \cdot \mathbf{1}_n), \quad \text{and}$$
$$\mu_{\text{new}} = \mathbb{E}[\xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n] = \mu_\xi + \boldsymbol{\gamma}_{n,\text{new}}^T \boldsymbol{\Gamma}_n^{-1}(\mathbf{y}_o - \mu_\xi \cdot \mathbf{1}_n),$$

the predictive variances are:

$$\sigma_{n+1}^2 = \text{Var}[Y'_{n+1}|\mathcal{Y}_n] = \sigma_\xi^2 - \boldsymbol{\gamma}_{n,n+1}^T \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_{n,n+1},$$
$$\sigma_{\text{new}}^2 = \text{Var}[\xi(\mathbf{x}_{\text{new}}|\mathcal{Y}_n] = \sigma_\xi^2 - \boldsymbol{\gamma}_{n,\text{new}}^T \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_{n,\text{new}}, \quad \text{and}$$
$$\rho = \rho_{\text{new}}(\mathbf{x}_{n+1}) = \text{Corr}[Y'_{n+1}, \xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n] = \sigma_\xi^2 R_{\boldsymbol{\theta}_\xi}(\mathbf{x}_{n+1}, \mathbf{x}_{\text{new}}) - \boldsymbol{\gamma}_{n,n+1}^T \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_{n,\text{new}}.$$

Here, $\boldsymbol{\gamma}_{n,n+1} = \sigma_\xi^2 \left[R_{\boldsymbol{\theta}_\xi}(\mathbf{x}_1, \mathbf{x}_{n+1}), \cdots, R_{\boldsymbol{\theta}_\xi}(\mathbf{x}_n, \mathbf{x}_{n+1})\right]^T$.

We then calculate the first two moments of truncated (bivariant) normal distribution:

$$\begin{aligned}
\mathbb{E}\left[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} \geq c\right] &= \int_{-\infty}^{\infty} y_{\text{new}} \times \psi_{Y_{\text{new}}|Y_{n+1} \geq c}(y_{\text{new}})\, dy_{\text{new}} \\
&= \frac{1}{1 - \Phi(z_c)} \int_c^{\infty} \psi_{Y_{n+1}}(y_{n+1})\, dy_{n+1} \int_{-\infty}^{\infty} y_{\text{new}} \psi_{Y_{\text{new}}|Y_{n+1}}(y_{\text{new}})\, dy_{\text{new}} \\
&= \mu_{\text{new}} + \rho \sigma_{\text{new}} \frac{\phi(z_c)}{1 - \Phi(z_c)}, \qquad \text{(B.4)}
\end{aligned}$$

where $z_c = (c - \mu_{n+1})/\sigma_{n+1}$ is the normalized upper censoring limit, $\psi_X(\cdot)$ is the probability density function (PDF) of random variable X, $\phi(\cdot)$ is the PDF of standard normal distribution, and $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. Similarly, we have:

$$\mathbb{E}\left[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} < c\right] = \mu_{\text{new}} - \rho \sigma_{\text{new}} \frac{\phi(z_c)}{\Phi(z_c)}, \quad \text{and} \qquad \text{(B.5)}$$

$$\text{Var}\left[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} \geq c\right] = \sigma_{\text{new}}^2 \left[1 + \rho^2 z_c \frac{\phi(z_c)}{1 - \Phi(z_c)} - \rho^2 \left(\frac{\phi(z_c)}{1 - \Phi(z_c)}\right)^2\right]. \qquad \text{(B.6)}$$

Furthermore, since the conditional variance of the joint normal distribution does not

depend on the value of Y_{n+1}, we have:

$$\mathbb{E}_{Y_{n+1}|Y'_{n+1}<c}[\text{Var}(\xi(\mathbf{x}_{\text{new}})|Y_{n+1})] = \text{Var}(\xi(\mathbf{x}_{\text{new}})|Y_{n+1}) = (1-\rho^2)\sigma^2_{\text{new}}. \tag{B.7}$$

Finally, plugging (B.4), (B.5), (B.6), and (B.7) back to (B.3), and using the fact that the probability of censoring $\lambda = \mathbb{P}[Y'_{n+1} \geq c|\mathcal{Y}_n] = 1 - \Phi(z_c)$, we have:

$$\text{CMSE}(\mathbf{x}_{n+1}, \mathbf{x}_{\text{new}}) = \sigma^2_{\text{new}} - \sigma^2_{\text{new}}\rho^2 \left[\Phi(z_c) - z_c\phi(z_c) + \frac{\phi^2(z_c)}{1-\Phi(z_c)}\right].$$

Therefore, with no censoring in training data, the proposed ICMSE criterion is:

$$\text{ICMSE}(\mathbf{x}_{n+1}) = \int_{[0,1]^p} \sigma^2_{\text{new}} - \sigma^2_{\text{new}}\rho^2 \left[\Phi(z_c) - z_c\phi(z_c) + \frac{\phi^2(z_c)}{1-\Phi(z_c)}\right] d\mathbf{x}_{\text{new}}. \tag{B.8}$$

B.1.3 Proof of Theorem 3

Consider a more general case *with* censoring in training data, i.e., $\mathcal{Y}_n = \{\mathbf{y}_o, \mathbf{y}'_c \geq \mathbf{c}\}$. It is important to note that, due to the existence of censoring data $\{\mathbf{y}'_c \geq \mathbf{c}\}$, the random variable $\xi(\mathbf{x}_{\text{new}})|\mathcal{Y}_n$ is *no longer* normally distributed. This, in turn, requires more cumbersome derivations than the earlier case without censoring in training data.

Using the conditional mean and variance expressions (2.3) and (2.4) for the *censored* GP, we have

$$\begin{aligned}
\mathbb{E}\left[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} \geq c\right] &= \mu_\xi + \boldsymbol{\gamma}^T_{n+1,\text{new}}\boldsymbol{\Gamma}^{-1}_{n+1}\left([\mathbf{y}_o, \hat{\mathbf{y}}_c, \hat{y}^{(>)}_{n+1}]^T - \mu_\xi \cdot \mathbf{1}_{n+1}\right), \\
\mathbb{E}\left[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} < c\right] &= \mu_\xi + \boldsymbol{\gamma}^T_{n+1,\text{new}}\boldsymbol{\Gamma}^{-1}_{n+1}\left([\mathbf{y}_o, \hat{\mathbf{y}}_c, \hat{y}^{(<)}_{n+1}]^T - \mu_\xi \cdot \mathbf{1}_{n+1}\right), \\
\text{Var}[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} \geq c] &= \sigma^2_\xi - \boldsymbol{\gamma}^T_{n+1,\text{new}}\boldsymbol{\Gamma}^{-1}_{n+1}\left(\boldsymbol{\Gamma}_{n+1} - \boldsymbol{\Sigma}_1\right)\boldsymbol{\Gamma}^{-1}_{n+1}\boldsymbol{\gamma}_{n+1,\text{new}}, \\
\mathbb{E}_{Y_{n+1}|Y'_{n+1}<c}[\text{Var}(\xi(\mathbf{x}_{\text{new}})|Y_{n+1})] &= \sigma^2_\xi - \boldsymbol{\gamma}^T_{n+1,\text{new}}\boldsymbol{\Gamma}^{-1}_{n+1}\left(\boldsymbol{\Gamma}_{n+1} - \hat{\boldsymbol{\Sigma}}\right)\boldsymbol{\Gamma}^{-1}_{n+1}\boldsymbol{\gamma}_{n+1,\text{new}},
\end{aligned} \tag{B.9}$$

where μ_ξ and σ^2_ξ are the mean and variance for the prior GP, respectively. Here, $\boldsymbol{\gamma}_{n+1,\text{new}} =$

$\sigma_\xi^2\big[R_{\boldsymbol{\theta}_\xi}(\mathbf{x}_1, \mathbf{x}_{\text{new}}), \cdots, R_{\boldsymbol{\theta}_\xi}(\mathbf{x}_{n+1}, \mathbf{x}_{\text{new}})\big]^T$ and $\boldsymbol{\Gamma}_{n+1} = \sigma_\xi^2[R_{\boldsymbol{\theta}_\xi}(\mathbf{x}_i, \mathbf{x}_j)]_{i=1\ j=1}^{n+1 n+1} + \sigma_\epsilon^2\mathbf{I}_{n+1}$. Furthermore, $\hat{y}_{n+1}^{(>)} = \mathbb{E}(Y'_{n+1}|\mathbf{y}_o, Y'_{n+1} \geq c)$ and $\hat{y}_{n+1}^{(<)} = \mathbb{E}(Y'_{n+1}|\mathbf{y}_o, Y'_{n+1} < c)$ are the expected response for the potential observation, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_1(\mathbf{x}_{n+1}) = \text{diag}(\mathbf{0}_{n_o}, \text{Var}([\mathbf{y}'_c, Y_{n+1}]|$ $\mathbf{y}_o, \mathbf{y}'_c \geq \mathbf{c}, Y_{n+1} = c))$, and $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}(\mathbf{x}_{n+1}) = \text{diag}\left(\mathbf{0}_{n_o}, \mathbb{E}_{Y_{n+1}|\mathbf{y}_o}[\text{Var}(\mathbf{y}'_c|Y_{n+1}, \mathbf{y}_o, \mathbf{y}'_c \geq \mathbf{c})], 0\right)$. Plugging in equation (B.9) back into (B.3), we have

$$\begin{aligned}\text{CMSE} = \sigma_{\text{new}}^2 &- \lambda(1-\lambda)\left(\boldsymbol{\gamma}_{n+1,\text{new}}^T\boldsymbol{\Gamma}_{n+1}^{-1}[\mathbf{0}_n, \hat{y}_{n+1}^{(>)} - \hat{y}_{n+1}^{(<)}]^T\right)^2 \\ &- (1-\lambda)\boldsymbol{\gamma}_{n+1,\text{new}}^T\boldsymbol{\Gamma}_{n+1}^{-1}\left(\boldsymbol{\Sigma}_1 - \hat{\boldsymbol{\Sigma}}\right)\boldsymbol{\Gamma}_{n+1}^{-1}\boldsymbol{\gamma}_{n+1,\text{new}}.\end{aligned}$$

Here, $\sigma_{\text{new}}^2 = \sigma_\xi^2 - \boldsymbol{\gamma}_{n,\text{new}}^T\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_{n,\text{new}} + \boldsymbol{\gamma}_{n,\text{new}}^T\boldsymbol{\Gamma}_n^{-1}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_{n,\text{new}}$ and the probability of censoring $\lambda = \mathbb{P}(\mathbf{y}_c \geq \mathbf{c}, Y'_{n+1} \geq c|\mathbf{y}_o)/\mathbb{P}(\mathbf{y}_c \geq \mathbf{c}|\mathbf{y}_o)$. (The computation of these orthant probabilities and moments of the truncated multivariate normal distribution will be discussed later in Appendix B.3.) Putting everything together, our ICMSE criterion has the following explicit form:

$$\text{ICMSE} = \int_{[0,1]^p} \sigma_{\text{new}}^2 - \boldsymbol{\gamma}_{n+1,\text{new}}^T\boldsymbol{\Gamma}_{n+1}^{-1}\mathbf{H}_c(\mathbf{x}_{n+1})\boldsymbol{\Gamma}_{n+1}^{-1}\boldsymbol{\gamma}_{n+1,\text{new}}\, d\mathbf{x}_{\text{new}} \tag{B.10}$$

where $\mathbf{H}_c(\mathbf{x}_{n+1}) = (1-\lambda)\left(\boldsymbol{\Sigma}_1 - \hat{\boldsymbol{\Sigma}}\right) + \lambda(1-\lambda)\text{diag}\left(\mathbf{0}_n, \hat{y}_{n+1}^{(>)}\hat{y}_{n+1}^{(<)}\right)$.

B.2 Multi-fidelity ICMSE design criterion

B.2.1 Proof of Theorem 4

For the multi-fidelity setting, the training data is $\{\mathbf{f}, \mathcal{Y}_n\}$. Using the conditional mean and variance expressions (2.16) and (2.17) for the *multi-fidelity* GP model, we have

$$\begin{aligned}
\mathbb{E}\left[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} \geq c\right] &= \mu_f + \boldsymbol{\gamma}_{n+1,\text{new}}^T \boldsymbol{\Gamma}_{n+1}^{-1} \left([\mathbf{f}, \mathbf{y}_o, \hat{\mathbf{y}}_c, \hat{y}_{n+1}^{(>)}]^T - \mu_f \cdot \mathbf{1}_{n+1}\right), \\
\mathbb{E}\left[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} < c\right] &= \mu_f + \boldsymbol{\gamma}_{n+1,\text{new}}^T \boldsymbol{\Gamma}_{n+1}^{-1} \left([\mathbf{f}, \mathbf{y}_o, \hat{\mathbf{y}}_c, \hat{y}_{n+1}^{(<)}]^T - \mu_f \cdot \mathbf{1}_{n+1}\right), \\
\text{Var}[\xi(\mathbf{x}_{\text{new}})|Y'_{n+1} \geq c] &= \sigma_f^2 + \sigma_\delta^2 - \boldsymbol{\gamma}_{n+1,\text{new}}^T \boldsymbol{\Gamma}_{n+1}^{-1} \left(\boldsymbol{\Gamma}_{n+1} - \boldsymbol{\Sigma}_1\right) \boldsymbol{\Gamma}_{n+1}^{-1} \boldsymbol{\gamma}_{n+1,\text{new}}, \\
\mathbb{E}_{Y_{n+1}|Y'_{n+1}<c}[\text{Var}(\xi(\mathbf{x}_{\text{new}})|Y_{n+1})] &= \sigma_f^2 + \sigma_\delta^2 - \boldsymbol{\gamma}_{n+1,\text{new}}^T \boldsymbol{\Gamma}_{n+1}^{-1} \left(\boldsymbol{\Gamma}_{n+1} - \hat{\boldsymbol{\Sigma}}\right) \boldsymbol{\Gamma}_{n+1}^{-1} \boldsymbol{\gamma}_{n+1,\text{new}}.
\end{aligned} \tag{B.11}$$

Though equations (B.11) appears to be quite similar to the equations (B.9), the notations in (B.11) are overloaded with the multi-fidelity expressions (2.16) and (2.17) for simplicity (see Section 2.3.1). Here, μ_f and σ_f^2 are the mean and variance of the GP $f(\cdot)$ modeling the computer experiment, $\boldsymbol{\gamma}_{n+1,\text{new}} = \sigma_f^2[R_{\boldsymbol{\theta}_f}(\mathbf{x}_i, \mathbf{x}_{\text{new}})]_{i=1}^{n+1} + \sigma_\delta^2[\mathbf{0}_{n-m}, R_{\boldsymbol{\theta}_\delta}(\mathbf{x}_i, \mathbf{x}_{\text{new}})]_{i=1}^{m+1}$, and $\boldsymbol{\Gamma}_{n+1} = \sigma_f^2[R_{\boldsymbol{\theta}_f}(\mathbf{x}_i, \mathbf{x}_j)]_{i=1\ j=1}^{n+1\ n+1} + \text{diag}\left(\mathbf{0}_{n-m}, \sigma_\delta^2[R_{\boldsymbol{\theta}_\delta}(\mathbf{x}_i, \mathbf{x}_j)]_{i=1\ j=1}^{m+1\ m+1} + \sigma_\epsilon^2 \mathbf{I}_{m+1}\right)$. Furthermore, $\hat{y}_{n+1}^{(>)} = \mathbb{E}(Y'_{n+1}|\mathbf{f}, \mathbf{y}_o, Y'_{n+1} \geq c)$ and $\hat{y}_{n+1}^{(<)} = \mathbb{E}(Y'_{n+1}|\mathbf{f}, \mathbf{y}_o, Y'_{n+1} < c)$ are the expected responses for the potential observation, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_1(\mathbf{x}_{n+1}) = \text{diag}(\mathbf{0}_{n-n_c}, \text{Var}(\mathbf{y}'_c \geq \mathbf{c}, Y_{n+1} = c|\mathbf{f}, \mathbf{y}_o))$, and $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}(\mathbf{x}_{n+1}) = \text{diag}(\mathbf{0}_{n-n_c}, \mathbb{E}_{Y_{n+1}|\mathbf{f},\mathbf{y}_o}[\text{Var}(\mathbf{y} \geq \mathbf{c}|Y_{n+1}, \mathbf{f}, \mathbf{y}_o, \mathbf{y}'_c \geq \mathbf{c})], 0)$.

Plugging in (B.11) to (B.3), we have the following explicit form ICMSE criterion:

$$\text{ICMSE}(\mathbf{x}_{n+1}) = \int_{[0,1]^p} \sigma_{\text{new}}^2 - \boldsymbol{\gamma}_{n+1,\text{new}}^T \boldsymbol{\Gamma}_{n+1}^{-1} \mathbf{H}_c(\mathbf{x}_{n+1}) \boldsymbol{\Gamma}_{n+1}^{-1} \boldsymbol{\gamma}_{n+1,\text{new}}\, d\mathbf{x}_{\text{new}}, \tag{B.12}$$

where, $\sigma_{\text{new}}^2 = \sigma_f^2 + \sigma_\delta^2 - \boldsymbol{\gamma}_{n,\text{new}}^T \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_{n,\text{new}} + \boldsymbol{\gamma}_{n,\text{new}}^T \boldsymbol{\Gamma}_n^{-1} \hat{\boldsymbol{\Sigma}} \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_{n,\text{new}}$, and

$$\mathbf{H}_c(\mathbf{x}_{n+1}) = (1-\lambda)\left(\boldsymbol{\Sigma}_1 - \hat{\boldsymbol{\Sigma}}\right) + \lambda(1-\lambda)\text{diag}\left(\mathbf{0}_n, \hat{y}_{n+1}^{(>)} \hat{y}_{n+1}^{(<)}\right). \tag{B.13}$$

Here, the probability of censoring $\lambda = \mathbb{P}(\mathbf{y}_c \geq \mathbf{c}, Y'_{n+1} \geq c|\mathbf{f}, \mathbf{y}_o)/\mathbb{P}(\mathbf{y}_c \geq \mathbf{c}|\mathbf{f}, \mathbf{y}_o)$.

B.2.2 Proof of Corollary 1

Note that in ICMSE criterion (2.18), $\boldsymbol{\gamma}_{n+1,\text{new}}$ is a function of both $\mathbf{x}_{n+1}$ and $\mathbf{x}_{\text{new}}$, $\boldsymbol{\Gamma}_{n+1}$ and $\mathbf{H}_c(\mathbf{x}_{n+1})$ are only function of $\mathbf{x}_{n+1}$, and σ^2_{new} is only a function of $\mathbf{x}_{\text{new}}$; it can therefore be further simplified as

$$\text{ICMSE}(\mathbf{x}_{n+1}) = \bar{\sigma}^2 - \text{tr}\left(\boldsymbol{\Gamma}_{n+1}^{-1}\mathbf{H}_c(\mathbf{x}_{n+1})\boldsymbol{\Gamma}_{n+1}^{-1}\boldsymbol{\Lambda}\right),$$

where $\bar{\sigma}^2 = \int \sigma^2_{\text{new}}\, d\mathbf{x}_{\text{new}}$ is a constant with respect to $\mathbf{x}_{n+1}$, tr(A) $= \sum_i A_{i,i}$ is the trace of matrix A, and $\boldsymbol{\Lambda} = \int \boldsymbol{\gamma}^T_{n+1,\text{new}}\boldsymbol{\gamma}_{n+1,\text{new}}\, d\mathbf{x}_{\text{new}}$. Assume the following product correlation structure:

$$R_{\boldsymbol{\theta}_f}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} R^{(l)}_{\boldsymbol{\theta}_f}(x_l, x'_l), \quad R_{\boldsymbol{\theta}_\delta}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} R^{(l)}_{\boldsymbol{\theta}_\delta}(x_l, x'_l), \tag{B.14}$$

and denote $\zeta^{(l)}(z, x) = R^{(l)}_{\boldsymbol{\theta}_f}(z, x) + \mathbb{1}_{\{i\geq(n-m)\}}R^{(l)}_{\boldsymbol{\theta}_\delta}(z, x)$. We can simplify the p-dimensional integral in $\boldsymbol{\Lambda}$ to a product of 1-dimensional integrals:

$$\Lambda_{ij} = \int_{[0,1]^p} \prod_{l=1}^{p} \zeta^{(l)}(x_{i,l}, x_l)\zeta^{(l)}(x_{j,l}, x_l)\, d\mathbf{x} = \prod_{l=1}^{p}\left[\int_0^1 \zeta^{(l)}(x_{i,l}, x_l)\zeta^{(l)}(x_{j,l}, x_l)\, dx_l\right], \tag{B.15}$$

where $i, j = 1, 2, \cdots, n+1$.

Furthermore, under the following product *Gaussian* correlation structure:

$$R_{\boldsymbol{\theta}_f}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} \theta_{f,\,l}^{4(x_i - x'_i)^2}, \quad R_{\boldsymbol{\theta}_\delta}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} \theta_{\delta,\,l}^{4(x_i - x'_i)^2}, \tag{B.16}$$

the 1-dimensional integrals of entries in $\boldsymbol{\Lambda}$ (B.15) can be reduced to integrals of exponential polynomial expressions, which have a closed form. Consider the following general

expression for an exponential polynomial:

$$\begin{aligned} G([a,x],[b,y]) &= \int_0^1 \exp\left[-a(x-z)^2\right] \exp\left[-b(y-z)^2\right] dz \\ &= \sqrt{\frac{\pi}{a+b}} \exp\left(\frac{(ax+by)^2}{a+b} - ax^2 - by^2\right) \\ &\times \left[\Phi\left(\sqrt{\frac{2}{a+b}}\,(a+b-ax-by)\right) - \Phi\left(-\sqrt{\frac{2}{a+b}}\,(ax+by)\right)\right], \end{aligned}$$

where $\Phi(\cdot)$ is the CDF for standard normal. Under (B.16), the entries of $\mathbf{\Lambda}$ (B.15) can be further simplified as:

$$\begin{aligned} \Lambda_{ij} =& \prod_{l=1}^{p} G([\tilde{\theta}_{f,l}, x_{i,l}], [\tilde{\theta}_{f,l}, x_{j,l}]) + \mathbb{1}_{\{i\geq(n-m)\}}\mathbb{1}_{\{j\geq(n-m)\}} \prod_{l=1}^{p} G([\tilde{\theta}_{\delta,l}, x_{i,l}], [\tilde{\theta}_{\delta,l}, x_{j,l}]) \\ &+\mathbb{1}_{\{i\geq(n-m)\}} \prod_{l=1}^{p} G([\tilde{\theta}_{f,l}, x_{i,l}], [\tilde{\theta}_{\delta,l}, x_{j,l}]) + \mathbb{1}_{\{j\geq(n-m)\}} \prod_{l=1}^{p} G([\tilde{\theta}_{\delta,l}, x_{i,l}], [\tilde{\theta}_{f,l}, x_{j,l}]), \end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_f = -4\log\boldsymbol{\theta}_f$ and $\tilde{\boldsymbol{\theta}}_\delta = -4\log\boldsymbol{\theta}_\delta$.

Note that the above simplification under product Gaussian correlations can also be used in the single-fidelity ICMSE criterion (2.11) as well, in which case $\Lambda_{ij} = \prod_{l=1}^{p} G([\tilde{\theta}_{\xi,l}, x_{i,l}], [\tilde{\theta}_{\xi,l}, x_{j,l}])$, with $\tilde{\boldsymbol{\theta}}_\xi = -4\log\boldsymbol{\theta}_\xi$.

B.3 Computational approximations

In practice, the computation of the matrix $\hat{\mathbf{\Sigma}}$ in the single-fidelity expression (B.9) or the multi-fidelity expression (B.11) can be quite time-consuming, since a closed-form expression is difficult to obtain for the expected variance term (the conditional distributions $[Y_{n+1}|\mathbf{y}_o]$ and $[Y_{n+1}|\mathbf{f}, \mathbf{y}_o]$ are non-Gaussian). For the single-fidelity setting, we found the following approximation to be useful for efficient computation:

$$\mathbb{E}_{Y_{n+1}<c|\mathcal{Y}_n}[\text{Var}(\mathbf{y}_c'|\mathbf{y}_o, \mathbf{y}_c' \geq \mathbf{c}, Y_{n+1})] \approx \text{Var}(\mathbf{y}_c'|\mathbf{y}_o, \mathbf{y}_c' \geq \mathbf{c}, Y_{n+1} = \hat{Y}_{n+1}),$$

where $\hat{Y}_{n+1} = \mathbb{E}(Y_{n+1}|\mathcal{Y}_n, Y_{n+1} < c)$. Similar simplification also applies for the multi-fidelity setting. This can be viewed as a plug-in estimate (with $Y_{n+1} = \hat{Y}_{n+1}$) which approximates the conditional mean expression on the left-hand side. The right-hand approximation can be efficiently computed via the truncated moments of a multivariate normal distribution, as implemented in the R package `tmvtnorm`.

APPENDIX C

APPENDIX FOR CHAPTER 3

C.1 Proof of Theorem 5

Since the generator $G(\cdot)$ is obtained by (3.3) with the training error $< \epsilon$, i.e.,

$$\mathcal{W}(\mathcal{X}, G_{\#}[\mathcal{U}]) = \inf_{\gamma} \int_{\mathbb{X}\times\mathbb{X}} \|x - G(u)\|_2 d\gamma(x, G(u)) = d < \epsilon. \tag{C.1}$$

This means we have obtained the transportation map $\gamma : \mathbb{X} \times \mathbb{X} \mapsto [0, 1]$, s.t.,

$$\mathbb{E}_{\gamma}[\|X - G(U)\|_2] = \int \|x - G(u)\|_2 d\gamma(x, G(u)) = d, \tag{C.2}$$

For any realization $x_i \in \mathbb{X}$ of the random variable $X \sim \mathcal{X}$, one can find a $u_i \in \mathbb{F}$ using the following optimization scheme:

$$u_i = \underset{u\in\mathbb{F}}{\operatorname{argmin}} \|x_i - G(u)\|_2, \tag{C.3}$$

We denote this as $u_i = h(x_i)$. If we denote the conditional measure of γ as $\gamma_{x_i} = \gamma | X = x_i$. Given $X = x_i$ and $u_i = h(x_i)$, clearly, we have,

$$\|x_i - G(u_i)\|_2 \leq \mathbb{E}_{U\sim\gamma_{x_i}} \|x_i - G(U)\|_2, \tag{C.4}$$

Furthermore, recall the dual formula of the Wasserstein distance:

$$\mathcal{W}(\mathcal{X}, G_{\#}[\mathcal{U}]) = \sup_{\|D(\cdot)\|_L \leq 1} \mathbb{E}_{x\sim\mathcal{X}}[D(x)] - \mathbb{E}_{u\sim\mathcal{U}}[D(G(u))] < \epsilon, \tag{C.5}$$

Specifically, if let the function $D(x) = h(x) - E(x)$, we have :

$$\|\mathbb{E}_{x\sim\mathcal{X}}[h(x) - E(G(h(x)))] - \mathbb{E}_{u\sim\mathcal{U}}[u - E(G(u))]\| < \epsilon, \tag{C.6}$$

With x_i, u_i and the Lipschitz-L continues assumption on $G(\cdot)$, we have:

$$\|G(E(x_i))-x_i\|_2 \leq \|G(E(x_i))-G(u_i)\|_2+\|G(u_i)-x_i\|_2 \leq L\|E(x_i)-u_i\|_2+\|G(u_i)-x_i\|_2 \tag{C.7}$$

Now, replace the realization x_i with the random variable X and take the expectation over $X \sim \mathcal{X}$,

$$\mathbb{E}_{X\sim\mathcal{X}}\|G(E(X)) - X\|_2 \leq L\mathbb{E}_{X\sim\mathcal{X}}\|E(X) - h(X)\|_2 + \mathbb{E}_{X\sim\mathcal{X}}\|G(h(X)) - X\|_2. \tag{C.8}$$

Considering the way we choose $u_i = h(X_i)$ and the inequality (C.4), for the second term above, we have

$$\mathbb{E}_{X\sim\mathcal{X}}\|X - G(h(X))\|_2 \leq \mathbb{E}_{X\sim\mathcal{X}}\mathbb{E}_{U\sim\gamma_X}\|X - G(U)\|_2 = \mathbb{E}_{\gamma}\|X - G(U)\|_2 \leq \epsilon \tag{C.9}$$

As for the first term, we have:

$$\mathbb{E}_{X\sim\mathcal{X}}\|E(X) - h(X)\|_2 \leq \mathbb{E}_{X\sim\mathcal{X}}\|E(X) - E(G(U))\|_2 + \mathbb{E}_{X\sim\mathcal{X}}\|h(X) - E(G(U))\|_2 \tag{C.10}$$

With the Lipschitz-L continues assumption on $E(\cdot)$:

$$\begin{aligned}\mathbb{E}_{X\sim\mathcal{X}}\|E(X) - h(X)\|_2 &\leq L\mathbb{E}_{X\sim\mathcal{X}}\|X - G(U)\|_2 + \mathbb{E}_{X\sim\mathcal{X}}\|h(X) - E(G(U))\|_2 \\ &\leq L\epsilon + \mathbb{E}_{X\sim\mathcal{X}}\|h(X) - E(G(U))\|_2 \end{aligned} \tag{C.11}$$

Note that $E(\cdot)$ is obtained by (3.5) with training error $\mathbb{E}_{U\sim\mathcal{U}}\left[\|E(G(U))-U\|_2\right]<\delta$. Recall Equation (C.6),

$$\mathbb{E}_{X\sim\mathcal{X}}\|U-E(G(U))\|_2\leq\mathbb{E}_{U\sim\mathcal{U}}\|U-E(G(U))\|_2+\epsilon\leq\delta+\epsilon \tag{C.12}$$

Finally, we have

$$\mathbb{E}_{X\sim\mathcal{X}}\|G(E(X))-X\|_2\leq L(L\epsilon+\delta+\epsilon)+\epsilon=(L^2+L+1)\epsilon+L\delta. \tag{C.13}$$

C.2 Proof of Theorem 6

we denote the target measure as $\mathcal{X}$ with its empirical representation as $\mathcal{X}_n$, while $\mathcal{X}'$ as measure obtained by proposed approach with its empirical representation as $\mathcal{X}'_m$.

(i) As the training data size approach infinity, $\mathcal{X}'_n\to\mathcal{X}_n$, this because using Pinsker's inequality,

$$\left|\sum_{x_i\in\mathcal{D}}\mathcal{I}(x_i>y)-\sum_{x'_i\in\mathcal{D}'}\mathcal{I}(x'_i>y)\right|<\sqrt{KL(\mathcal{X}'_n||\mathcal{X}_n)}. \tag{C.14}$$

where $KL(\cdot||\cdot)$ is the K-L divergence of two distribution. From [118], we know the $KL(\mathcal{X}'_n||\mathcal{X}_n)\to 0$ as $n\to\infty$, i.e., existing a small $\epsilon>0$, for any y,

$$\left|\sum_{x_i\in\mathcal{D}}\mathcal{I}(x_i>y)-\sum_{x'_i\in\mathcal{D}'}\mathcal{I}(x'_i>y)\right|<\frac{\epsilon}{3}. \tag{C.15}$$

For more discussion and justification, please refer to [118].

(ii) As the data size approach infinity, $\mathcal{X}_n\to\mathcal{X}$. Since the training dataset $\mathcal{D}=\{x_i\}_{i=1}^n$ is sampled from the target measure $\mathcal{X}$, its empirical cumulative distribution function (CDF) converges to target CDF $F_{\mathcal{X}}$, i.e., for any y,

$$\left|\sum_{x_i\in\mathcal{D}}\mathcal{I}(x_i>y)-F_{\mathcal{X}}(y)\right|<\frac{\epsilon}{3}. \tag{C.16}$$

(iii) Similar to (ii), as the data size approach infinity, $\left|\sum_{x_i' \in \mathcal{D}'} \mathcal{I}(x_i' > y) - F_{\S'}(y)\right| < \epsilon/3$.

We have the difference in the obtained CDF and the target CDF $\left|F_{\mathcal{X}}(y) - F_{\S'}(y)\right|$:

$$\left|F_{\mathcal{X}}(y) - \sum_{x_i \in \mathcal{D}} \mathcal{I}(x_i > y) + \sum_{x_i \in \mathcal{D}} \mathcal{I}(x_i > y) - \sum_{x_i' \in \mathcal{D}'} \mathcal{I}(x_i' > y) + \sum_{x_i' \in \mathcal{D}'} \mathcal{I}(x_i' > y) - F_{\S'}(y)\right|. \tag{C.17}$$

Combining (i), (ii) and (iii), we know as the training data size large enough, with any y,

$$\begin{aligned}\left|F_{\mathcal{X}}(y) - F_{\S'}(y)\right| \leq & \left|F_{\mathcal{X}}(y) - \sum_{x_i \in \mathcal{D}} \mathcal{I}(x_i > y)\right| + \left|\sum_{x_i \in \mathcal{D}} \mathcal{I}(x_i > y) - \sum_{x_i' \in \mathcal{D}'} \mathcal{I}(x_i' > y)\right| \\ & + \left|\sum_{x_i' \in \mathcal{D}'} \mathcal{I}(x_i' > y) - F_{\S'}(y)\right| < \epsilon. \end{aligned} \tag{C.18}$$

i.e., as the training data size approach infinity, $\mathcal{X}' \to \mathcal{X}$ in distribution.

C.3 Proof of Theorem 7

Note that the objective function in (3.10) is not a energy distance. However, we have

$$\begin{aligned} & \underset{\{f_1', \cdots, f_m'\}}{\operatorname{argmin}} \sum_{i=1}^{n} \mathbb{E}_{\gamma \sim \mu_h} \|f_i' - \gamma\|_2 - \frac{1}{2(m+n)} \sum_{i=1}^{m+n} \sum_{j=1}^{m+n} \|f_i' - f_j'\|_2 \\ = & \underset{\{f_1', \cdots, f_m'\}}{\operatorname{argmin}} \sum_{i=1}^{m+n} \mathbb{E}_{\gamma \sim \mu_h} \|f_i' - \gamma\|_2 + \sum_{i=m+1}^{m+n} \mathbb{E}_{\gamma \sim \mu_h} \|f_i' - \gamma\|_2 - \frac{1}{2(m+n)} \sum_{i=1}^{m+n} \sum_{j=1}^{m+n} \|f_i' - f_j'\|_2 \\ = & \underset{\{f_1', \cdots, f_m'\}}{\operatorname{argmin}} \; dist(\mathcal{F}_{n+m}', \mu_h) \end{aligned} \tag{C.19}$$

Here, m is the number of virtual points and n is the number of actual points. Here, we set $f_i' = f_i$ for actual points with indies $i = m+1, \cdots, m+n$ and let $\mathcal{F}_{n+m}'$ be the empirical measure for features $\{f_i'\}_{i=1}^{n+m}$. Note that the minimizing is only on the virtual dataset, i.e., with subscripts $1, \cdots, m$.

Note that for energy distance

$$dist(\mathcal{F}'_{n+m}, \mu_h) \leq dist(\mathcal{F}'_{n+m}, \mathcal{F}'_m) + dist(\mathcal{F}'_m, \mu_h) \tag{C.20}$$

As $m \to \infty$, the second term in (C.20) approaches to zero, and equivalently we are minimizing $dist(\mathcal{F}'_m, \mu_h)$. Following [123], we have $\mathcal{F}'_m \to \mu_h$ in distribution.

Furthermore, with the continuity condition on the $G(\cdot)$, we have

$$\mathcal{X}'_m \to \mu_H \tag{C.21}$$

directly following the continuous mapping theorem.

C.4 Details of the implementation

We explain the implementation details here.

Four-fold cross validation. A four-fold cross validation strategy is used in the aortic stenosis application. We have 168 data in total. Three quarters of the data ($168 \times 75\% \times 10 = 1260$) after rotation augmentation is used as the training set, while the remaining quarter ($168 \times 25\% = 42$) will be the testing set. Since the architecture of the classifier is pre-defined, and there are no hyperparameters that need to be tuned, the validation set is not needed.

Training GIN. For the synthetic dataset, the generator $G(\cdot)$ adapts 5-layer vanilla NN with 512, 512, 1024, 1024, 1024 hidden nodes in each hidden layer, respectively, and ReLu activation. The discriminator $D(\cdot)$ also adapts 5-layer vanilla NN with 1024, 1024, 1024, 512, 512 hidden nodes, and ReLu activation. As for the encoder $E(\cdot)$, it has 10 convolutional layers with 128, 256, 256, 512, 512, 1024, 1024, 1024, 512, 256 hidden nodes in each hidden layer, respectively, leaky ReLu activation and batch normalization. In our implementation, we train the GIN for 2000 epochs, with a constant learning rate of $1e-5$. For the aortic stenosis applications, the architecture and the training strategy are similar, except that the

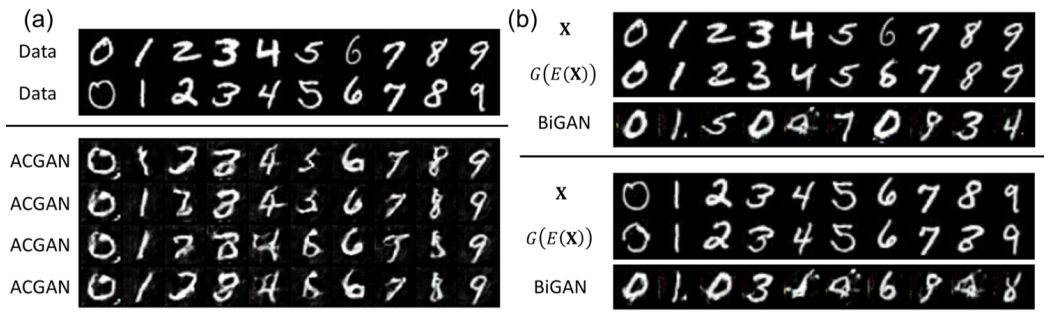

Figure C.1: *Qualitative results for GIN training using MNIST, including the training set data* $\mathbf{X}$ *of different classes, generated samples via ACGAN, our reconstructions* $G(E(\mathbf{X}))$ *and reconstructions via BiGAN.*

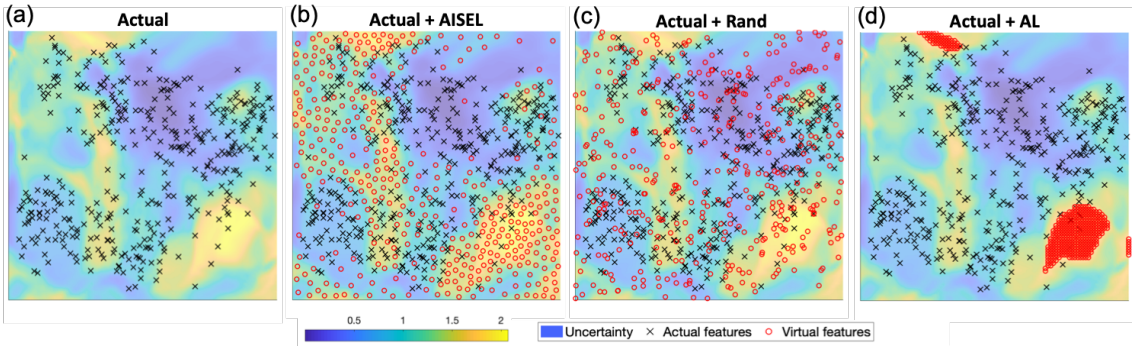

Figure C.2: *A comparison of the selected features by our AISEL method, the random sampling method, and the active learning method on the MNIST dataset, with the uncertainty measure* (3.7) *as background.*

numbers of the hidden nodes and training epochs are doubled.

Training native and improved models. For the toy computer vision datasets, both the native model $C(\cdot)$ and the improved model $C^*(\cdot)$ have three convolutional layers with 32, 64 and 64 hidden nodes, respectively. Leaky ReLu activation and batch normalization are also included in each layer. After the convolutional layers, two fully connected layers with 512 and 64 hidden nodes are used, respectively, with ReLu activation and batch normalization. Cross-entropy loss is used. In our implementation, we train the CNN for 80 epochs. The initial learning rate is $1e-4$, with decay to a half every 20 epochs. We select the above for the best empirical performance in preliminary experiments. For the aortic stenosis application, both $C(\cdot)$ and $C^*(\cdot)$ have the similar three convolutional layers with leaky ReLu activation and batch normalization. The three convolutional layers have 32, 64 and 128

hidden nodes, respectively. After the convolutional layers, three fully connected layers with 512, 128 and 32 hidden nodes are used, respectively, with ReLu activation and batch normalization. Cross-entropy loss and the same decay of learning rate is used. Note that the complexity for both $C(\cdot)$ and $C^*(\cdot)$ is low, especially compared to the encoder $E(\cdot)$. This is mainly because of the difference in the classification task for $C(\cdot)$ and regression task for $E(\cdot)$.

C.5 Toy MNIST experiments

We conduct the same experiments on the MNIST dataset as the Fashion dataset in Section 5.1; the setup and the implementation details are the same as that in the Fashion experiments. Figure C.1 shows the visual comparison of the proposed GIN and baselines. We see that in Figure C.1 (b), our GIN can generate shape images, with visual superior reconstruction performance than the BiGAN. As for ACGAN (see Figure C.1 (a)), we observe that the performance is not as good as the proposed GIN.

The final classification performance is already shown in Table C.1. Our AISEL method achieves predictive accuracy of 91.2%, a $91.2\% - 88.2\% = 3\%$ improvement compared to the native model. Meanwhile, our method outperforms the baselines, e.g., transfer learning, ACGAN-based method, and active learning. As for the GIN-based random augmentation, our method achieves (i) better performance when the same amount of virtual data (400) is used and (ii) similar performance when 5000 data is used in the baseline. This superior performance is again contributing to the exploration and exploitation of the feature space (see Figure C.2).

C.6 Balancing the label distribution

The proposed sampling method can also be used to balance the label distribution. Note that the uncertainty measure defined in (3.7) is not normalized. In order to balance the data, we

Table C.1: *A comparison of F1 score, area under the receiver operating characteristic curve (AUC), and classification accuracy of the native model and different improved models, under the imbalanced training dataset.*

	Native (300)	**Undersampling**	**Oversampling**
F1 score	**0.9448**	0.9589	0.9624
AUC	**0.9782**	0.9839	0.9846
Accuracy	**94.60%**	95.90%	96.25%
	Random (+600)	**AISEL (+300)**	**AISEL (+600)**
F1 score	0.9728	**0.9734**	**0.9801**
AUC	0.9905	**0.9937**	**0.9964**
Accuracy	96.95%	**97.05%**	**98.00%**

modify the uncertainty measure as

$$h_b(f_0) = \frac{h(f_0)}{\int h(f)\mathcal{I}\left[c(f) = c(f_0)\right]\ df_0}, \tag{C.22}$$

where notation $c(f) = \operatorname{argmax} C(G(f))$ denotes the label (rather than the predictive probability) of the native model, $\mathcal{I}[\cdot]$ is the indicator function. The denominator normalizes the density with respect to different classes, and therefore balances the label distribution.

We then apply this to learn a classification model from the truncated and imbalanced Fashion dataset. For the sake of simplicity, we consider two-class classification (i.e., using the two classes "Top" and "Coat"). The training dataset is designed to be both small ($n = 300$) and imbalanced (size of classes, $11:1$). We then applied the proposed AISEL framework as discussed in Section 5.1.1 using the same setup as discussed in Appendix C.4. Table C.1 lists the performance of the proposed AISEL method (with balancing via (C.22)), random sampling (without balancing), and the standard baselines of under-sampling and over-sampling. Compared with the native model, the two sampling strategies improve predictive performance marginally. In contrast, the improvement in predictive accuracy using the proposed AISEL method is around 2.5% and 3.5% with 300 and 600 virtual data points, respectively. Meanwhile, the proposed AISEL method mitigates the imbalance of the training dataset with improvements of at least 0.028 and 0.015 in the F1 score and AUC,

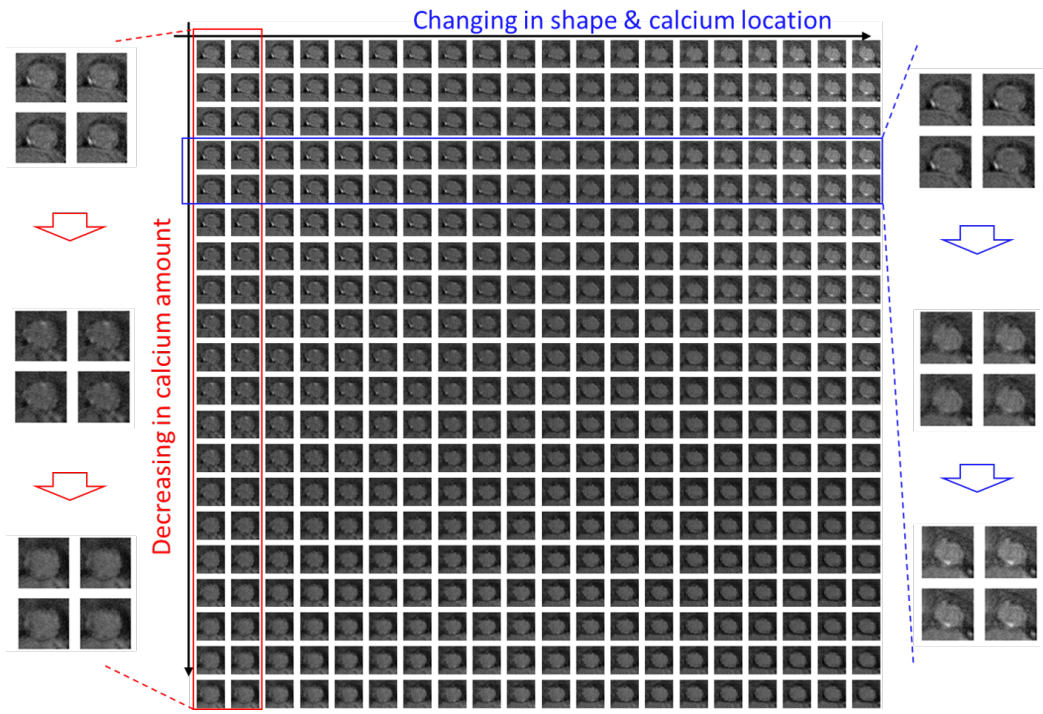

Figure C.3: *Qualitative visualization of 2D cross-section of feature space with the generated virtual images on the grid of feature space. The pathophysiological meaning of both axes is visualized in the left and right sides, respectively.*

respectively. Furthermore, compared to the randomly generated virtual dataset of size 600, the proposed AISEL method (with balancing) achieves noticeable improvements, even with a smaller data size of 300.

C.7 More on aortic stenosis application

Due to the limited space, Figure 3.7 in Section 5.2.2 only shows the partial 2D cross-section of the feature space. Figure C.3 visualizes the whole and enlarged 2D cross-section feature space. The two axes of the 2D cross-section shown in Figure C.3 have pathophysiological meaning. As shown in the red box (enlarged images on the left side), the vertical axis can be interpreted as the change of the calcification (i.e., the regions of high intensity in the CT images) amount. As shown in the blue box (enlarged images on the right side), the horizontal axis can be interpreted as the change of valve shape and the calcification location.

In order to better visualize the sampled AISEL dataset and demonstrate how it helps

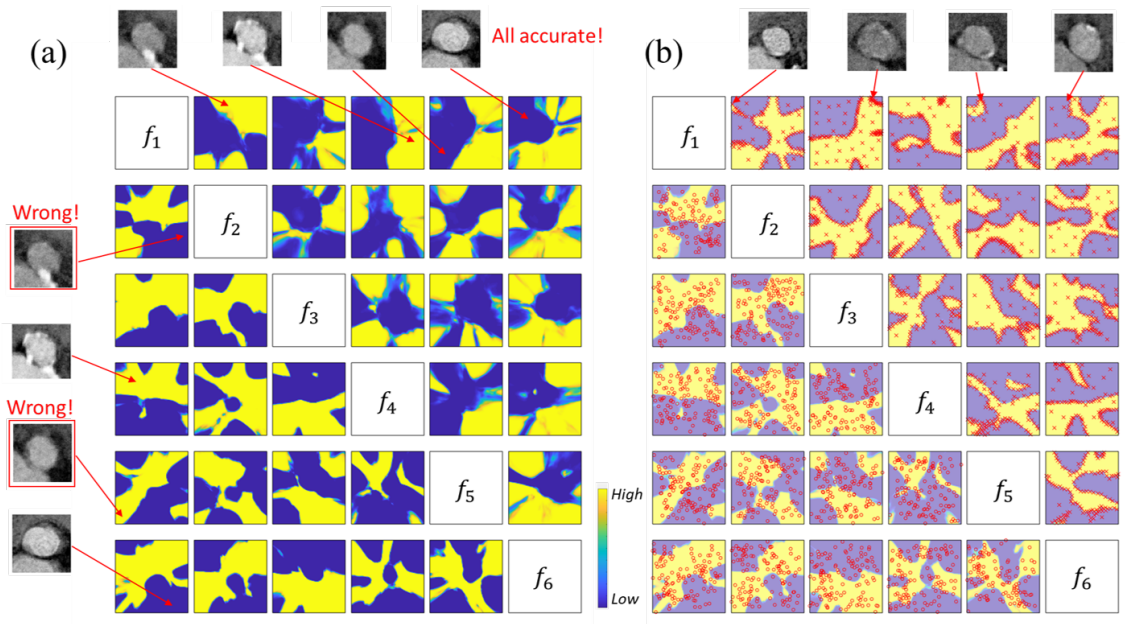

Figure C.4: *(a) A compression of the classification model of the native model (bottom left) and the improved model (top right) via proposed method on the 6D feature space $\mathbb{F}$. Four testing images are show and the native model can only correctly classify two of them, while the improved model can correctly classify all of them. (b) A compression of the actual patients (bottom left) and the selected features of our AISEL dataset (top right) in the feature space. Four examples of the generated virtual patients are also shown.*

in improving the classification accuracy, we conduct the experiments in the Section 3.5.2 with the dimension of the feature space $r = 6$, i.e., $\mathbb{F} = [-1, 1]^6$. Furthermore in the two-class classification problem, we use $C(G(\cdot)) : \mathbb{F} \mapsto [0, 1]$, with low value ~ 0 denoting the low calcification situation and high value ~ 1 for high calcification situation. The prediction contour $C(G(\cdot))$ of the model learned by the first three folds of the training data (the remaining fold is for testing) is shown in the lower-left half of Figure C.4 (a). Every small figure visualizes a 2D subspace of $\mathbb{F}$ with the remaining features set to be zero. Note that all combinations of the two features (totally 15 for six features) are shown in the lower-left half of Figure C.4 (a). Yellow means high calcification (i.e., $C(G(f)) = 1$), while blue means low calcification (i.e., $C(G(f)) = 0$). Meanwhile, four testing valves are shown in the left side of Figure C.4 (a). The obtained native model $C(\cdot)$ only accurately classifies two of them, which indicates the poor performance of the native model.

Designed AISEL data are shown in upper left region of Figure C.4 (b). Note that for

every figure, only 10% of the feature of AISEL dataset closest to the 2D cross-section plane are included for better visualization purpose. Most of AISEL features, as expected, are located on the boundary of the prediction contour of the native model, while the rest features are uniformly spread over the whole space. This again shows the exploration and exportation properties of the proposed AISEL method. Four examples of the selected virtual images in the AISEL dataset are also visualized on the top, with an arrow pointing their features in the feature space. Visually, they are indeed confusing for predicting the calcification amount. After physical labeling by a radiologist, they will help improve the classifier. As a comparison, the actual images (totally, 126) projected in every 2D cross-section are shown in the lower right region of Figure C.4 (b). Note that the actual images are randomly distributed in the whole 6D space with no apparent pattern.

The prediction contour $C^*(G(\cdot))$ of the improved classifier $C^*(\cdot)$ using our AISEL method is shown in the top left half of Figure C.4 (a). We can see the finer structure is learned indicating a more sophisticated model is obtained. Meanwhile, the four characteristic images tested by the native model is also tested by the $C^*(\cdot)$. The classification of all four is accurate, showing a noticeable improvement in the prediction accuracy.

REFERENCES

[1] F. Rengier, A. Mehndiratta, H. Von Tengg-Kobligk, C. M. Zechmann, R. Unterhinninghofen, H.-U. Kauczor, and F. L. Giesel, "3D printing based on imaging data: Review of medical applications", *International Journal of Computer Assisted Radiology and Surgery*, vol. 5, no. 4, pp. 335–341, 2010.

[2] J. Chen, Y. Xie, K. Wang, Z. H. Wang, G. Lahoti, C. Zhang, M. A. Vannan, B. Wang, and Z. Qian, "Generative invertible networks (GIN): Pathophysiology-interpretable feature mapping and virtual patient generation", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2018, pp. 537–545.

[3] Z. Qian, K. Wang, S. Liu, X. Zhou, V. Rajagopal, C. Meduri, J. R. Kauten, Y.-H. Chang, C. Wu, and C. Zhang, "Quantitative prediction of paravalvular leak in transcatheter aortic valve replacement based on tissue-mimicking 3D printing", *JACC: Cardiovascular Imaging*, vol. 10, no. 7, pp. 719–731, 2017.

[4] M. L. Raghavan, M. W. Webster, and D. A. Vorp, "Ex vivo biomechanical behavior of abdominal aortic aneurysm: Assessment using a new mathematical model", *Annals of Biomedical Engineering*, vol. 24, no. 5, pp. 573–582, 1996.

[5] K. Wang, Y. Zhao, Y.-H. Chang, Z. Qian, C. Zhang, B. Wang, M. A. Vannan, and M.-J. Wang, "Controlling the mechanical behavior of dual-material 3D printed meta-materials for patient-specific tissue-mimicking phantoms", *Materials & Design*, vol. 90, pp. 704–712, 2016.

[6] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*. Springer Science & Business Media, 2013.

[7] G. Mathéron, "Principles of geostatistics", *Economic Geology*, vol. 58, no. 8, pp. 1246–1266, 1963.

[8] M. Bayarri, J. Berger, J. Cafeo, G. Garcia-Donato, F. Liu, J. Palomo, R. Parthasarathy, R. Paulo, J. Sacks, and D. Walsh, "Computer model validation with functional output", *The Annals of Statistics*, vol. 35, no. 5, pp. 1874–1906, 2007.

[9] D. Higdon, J. Gattiker, B. Williams, and M. Rightley, "Computer model calibration using high-dimensional output", *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 570–583, 2008.

[10] S. Mak, C.-L. Sung, X. Wang, S. T. Yeh, Y. H. Chang, V. R. Joseph, V. Yang, and C. J. Wu, "An efficient surrogate model for emulation and physics extraction of

large eddy simulations", *Journal of the American Statistical Association*, pp. 1–14, 2018.

[11] S. Guillas, A. Sarri, S. J. Day, X. Liu, and F. Dias, "Functional emulation of high resolution tsunami modelling over cascadia", *The Annals of Applied Statistics*, vol. 12, no. 4, pp. 2023–2053, 2018.

[12] A. Stein and L. Corsten, "Universal kriging and cokriging as a regression procedure", *Biometrics*, vol. 47, no. 2, pp. 575–587, 1991.

[13] S. Banerjee, B. P. Carlin, and A. E. Gelfand, *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, 2014.

[14] H. Mohammadi, P. Challenor, and M. Goodfellow, "Emulating dynamic non-linear simulators using Gaussian processes", *Computational Statistics & Data Analysis*, vol. 139, pp. 178–196, 2019.

[15] J. Ramsay, *Functional Data Analysis*. Springer Series in Statistics, NY., 2005.

[16] J. Fan and W. Zhang, "Statistical methods with varying coefficient models", *Statistics and its Interface*, vol. 1, no. 1, p. 179, 2008.

[17] N. Malfait and J. O. Ramsay, "The historical functional linear model", *Canadian Journal of Statistics*, vol. 31, no. 2, pp. 115–128, 2003.

[18] M. D. Morris, "Gaussian surrogates for computer models with time-varying inputs and outputs", *Technometrics*, vol. 54, no. 1, pp. 42–50, 2012.

[19] T. Muehlenstaedt, J. Fruth, and O. Roustant, "Computer experiments with functional inputs and scalar outputs by a norm-based approach", *Statistics and Computing*, vol. 27, no. 4, pp. 1083–1097, 2017.

[20] B. Wang and A. Xu, "Gaussian process methods for nonparametric functional regression with mixed predictors", *Computational Statistics & Data Analysis*, vol. 131, pp. 80–90, 2019.

[21] L. E. Malvern, *Introduction to the Mechanics of a Continuous Medium*, Monograph. 1969.

[22] R. Hill, *The Mathematical Theory of Plasticity*. Oxford University Press, 1998, vol. 11.

[23] O. C. Zienkiewicz, R. L. Taylor, O. C. Zienkiewicz, and R. L. Taylor, *The Finite Element Method*. McGraw-Hill London, 1977, vol. 36.

[24] J. Chen, K. Wang, C. Zhang, and B. Wang, "An efficient statistical approach to design 3D-printed metamaterials for mimicking mechanical properties of soft biological tissues", *Additive Manufacturing*, vol. 24, pp. 341–352, 2018.

[25] V. R. Joseph, E. Gul, and S. Ba, "Maximum projection designs for computer experiments", *Biometrika*, vol. 102, no. 2, pp. 371–380, 2015.

[26] I. M. Sobol', "On the distribution of points in a cube and the approximate evaluation of integrals", *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, vol. 7, no. 4, pp. 784–802, 1967.

[27] R. N. Bracewell and R. N. Bracewell, *The Fourier Transform and its Applications*. McGraw-Hill New York, 1986, vol. 31999.

[28] X. Liu and S. Guillas, "Dimension reduction for Gaussian process emulation: An application to the influence of bathymetry on tsunami heights", *SIAM/ASA Journal on Uncertainty Quantification*, vol. 5, no. 1, pp. 787–812, 2017.

[29] J. Chen, Z. Liu, K. Wang, C. Jiang, C. Zhang, and B. Wang, "A calibration-free method for biosensing in cell manufacturing", *IISE Transactions*, pp. 1–39, 2020.

[30] T. Park and G. Casella, "The Bayesian lasso", *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[31] H. Ishwaran and J. S. Rao, "Spike and slab variable selection: Frequentist and bayesian strategies", *The Annals of Statistics*, vol. 33, no. 2, pp. 730–773, 2005.

[32] H. Wang, "Bayesian graphical lasso models and efficient posterior computation", *Bayesian Analysis*, vol. 7, no. 4, pp. 867–886, 2012.

[33] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso", *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[34] H. V. Sorensen, D. Jones, M. Heideman, and C. Burrus, "Real-valued fast Fourier transform algorithms", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 6, pp. 849–863, 1987.

[35] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC, 1995.

[36] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, 10. Springer Series in Statistics, NY., 2001, vol. 1.

[37] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization", *Mathematical Programming*, vol. 45, no. 1-3, pp. 503–528, 1989.

[38] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization", *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, 2001.

[39] X. Wang and J. O. Berger, "Estimating shape constrained functions using gaussian processes", *SIAM/ASA Journal on Uncertainty Quantification*, vol. 4, no. 1, pp. 1–25, 2016.

[40] R. B. Martin, D. B. Burr, and N. A. Sharkey, *Skeletal Tissue Mechanics*. Springer, 1998, vol. 190.

[41] L. Hockaday, K. Kang, N. Colangelo, P. Cheung, B. Duan, E. Malone, J. Wu, L. Girardi, L. Bonassar, H. Lipson, C. Chu, and J. Butcher, "Rapid 3D printing of anatomically accurate and mechanically heterogeneous aortic valve hydrogel scaffolds", *Biofabrication*, vol. 4, no. 3, p. 035 005, 2012.

[42] J. Chen, Y. Xie, K. Wang, C. Zhang, M. A. Vannan, B. Wang, and Z. Qian, "Active image synthesis for efficient labeling", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[43] R. Brooks, "On the loss of information through censoring", *Biometrika*, vol. 69, no. 1, pp. 137–144, 1982.

[44] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, "Design and analysis of computer experiments", *Statistical Science*, vol. 4, no. 4, pp. 409–423, 1989.

[45] I. Gibson, D. W. Rosen, and B. Stucker, *Additive Manufacturing Technologies*. Springer, 2014, vol. 17.

[46] K. Wang, C. Wu, Z. Qian, C. Zhang, B. Wang, and M. A. Vannan, "Dual-material 3D printed metamaterials with tunable mechanical properties for patient-specific tissue-mimicking phantoms", *Additive Manufacturing*, vol. 12, pp. 31–37, 2016.

[47] J. Chen, S. Mak, V. R. Joseph, and C. Zhang, "Function-on-function kriging, with applications to three-dimensional printing of aortic tissues", *Technometrics*, pp. 1–12, 2020.

[48] M. Quirk and J. Serda, *Semiconductor Manufacturing Technology*. Prentice Hall Upper Saddle River, NJ, 2001, vol. 1.

[49] R. Jin, C.-J. Chang, and J. Shi, "Sequential measurement strategy for wafer geometric profile estimation", *IIE Transactions*, vol. 44, no. 1, pp. 1–12, 2012.

[50] R. Singh, M. Fakhruddin, and K. Poole, "Rapid photothermal processing as a semiconductor manufacturing technology for the 21st century", *Applied Surface Science*, vol. 168, no. 1-4, pp. 198–203, 2000.

[51] M. Van Gurp and J. Palmen, "Time-temperature superposition for polymeric blends", *Rheology Bulletin*, vol. 67, no. 1, pp. 5–8, 1998.

[52] A. Feteira, "Negative temperature coefficient resistance (NTCR) ceramic thermistors: An industrial perspective", *Journal of the American Ceramic Society*, vol. 92, no. 5, pp. 967–983, 2009.

[53] C. G. Kaufman, D. Bingham, S. Habib, K. Heitmann, and J. A. Frieman, "Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology", *The Annals of Applied Statistics*, vol. 5, no. 4, pp. 2470–2492, 2011.

[54] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*. Springer Science & Business Media, 2018.

[55] B. Ankenman, B. L. Nelson, and J. Staum, "Stochastic kriging for simulation metamodeling", *Operations Research*, vol. 58, no. 2, pp. 371–382, 2010.

[56] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models", *Journal of the Royal Statistical Society: Series B*, vol. 63, no. 3, pp. 425–464, 2001.

[57] N. Henkenjohann, R. Göbel, M. Kleiner, and J. Kunert, "An adaptive sequential procedure for efficient optimization of the sheet metal spinning process", *Quality and Reliability Engineering International*, vol. 21, no. 5, pp. 439–455, 2005.

[58] P. Groot, P. Lucas, A. Cano, M. Gómez-Olmedo, and T. Nielsen, "Gaussian process regression with censored data using expectation propagation", in *Cano, A.; Gómez-Olmedo, M.; Nielsen, TD (ed.), PGM 2012: Proceedings of the Sixth European Workshop on Probabilistic Graphical Models, PGM'12 Granada, Spain September 19-21, 2012*, Granada: DECSAI, 2012, pp. 115–122.

[59] S. Da Veiga and A. Marrel, "Gaussian process modeling with inequality constraints", in *Annales de la Faculté des sciences de Toulouse: Mathématiques*, vol. 21, 2012, pp. 529–555.

[60] A. F. López-Lopera, F. Bachoc, N. Durrande, and O. Roustant, "Finite-dimensional Gaussian approximation with linear inequality constraints", *SIAM/ASA Journal on Uncertainty Quantification*, vol. 6, no. 3, pp. 1224–1255, 2018.

[61] L. Ding, S. Mak, and C. F. J. Wu, "Bdrygp: A new Gaussian process model for incorporating boundary information", *arXiv preprint arXiv:1908.08868*, 2019.

[62] F. Cao, S. Ba, W. A. Brenneman, and V. R. Joseph, "Model calibration with censored data", *Technometrics*, vol. 60, no. 2, pp. 255–262, 2018.

[63] D. M. Borth, "Optimal experimental designs for (possibly) censored data", *Chemometrics and Intelligent Laboratory Systems*, vol. 32, no. 1, pp. 25–35, 1996.

[64] E. M. Monroe and R. Pan, "Experimental design considerations for accelerated life tests with nonlinear constraints and censoring", *Journal of Quality Technology*, vol. 40, no. 4, pp. 355–367, 2008.

[65] M. E. Johnson, L. M. Moore, and D. Ylvisaker, "Minimax and maximin distance designs", *Journal of Statistical Planning and Inference*, vol. 26, no. 2, pp. 131–148, 1990.

[66] M. D. Morris and T. J. Mitchell, "Exploratory designs for computational experiments", *Journal of Statistical Planning and Inference*, vol. 43, no. 3, pp. 381–402, 1995.

[67] M. C. Shewry and H. P. Wynn, "Maximum entropy sampling", *Journal of Applied Statistics*, vol. 14, no. 2, pp. 165–170, 1987.

[68] C. Q. Lam, "Sequential adaptive designs in computer experiments for response surface model fit", Ph.D. dissertation, The Ohio State University, 2008.

[69] S. Xiong, P. Z. Qian, and C. F. J. Wu, "Sequential design and analysis of high-accuracy and low-accuracy computer codes", *Technometrics*, vol. 55, no. 1, pp. 37–46, 2013.

[70] R.-B. Chen, W. Wang, and C. F. J. Wu, "Sequential designs based on Bayesian uncertainty quantification in sparse representation surrogate modeling", *Technometrics*, vol. 59, no. 2, pp. 139–152, 2017.

[71] J. Bect, F. Bachoc, D. Ginsbourger, *et al.*, "A supermartingale approach to Gaussian process based sequential design of experiments", *Bernoulli*, vol. 25, no. 4A, pp. 2883–2919, 2019.

[72] M. Binois, J. Huang, R. B. Gramacy, and M. Ludkovski, "Replication or exploration? Sequential design for stochastic simulation experiments", *Technometrics*, vol. 61, no. 1, pp. 7–23, 2019.

[73] R. B. Gramacy and D. W. Apley, "Local Gaussian process approximation for large computer experiments", *Journal of Computational and Graphical Statistics*, vol. 24, no. 2, pp. 561–578, 2015.

[74] J. Sacks, S. B. Schiller, and W. J. Welch, "Designs for computer experiments", *Technometrics*, vol. 31, no. 1, pp. 41–47, 1989.

[75] J. Ypma, H. Borchers, and D. Eddelbuettel, "Nloptr: R interface to nlopt", *R Journal*, vol. 1, no. 4, 2014.

[76] J. A. Nelder and R. Mead, "A simplex method for function minimization", *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[77] S. Wilhelm and B. G. Manjunath, "Tmvtnorm: A package for the truncated multivariate normal distribution", *R Journal*, vol. 2, no. 1, pp. 25–29, 2010.

[78] V. R. Joseph, "Rejoinder", *Quality Engineering*, vol. 28, no. 1, pp. 42–44, 2016. eprint: https://doi.org/10.1080/08982112.2015.1100452.

[79] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation", *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[80] Thermo Electric Company, *Wafer sensors*, http://www.te-direct.com/products/silicon-wafers/, 2010.

[81] E. J. Dickinson, H. Ekström, and E. Fontes, "COMSOL Multiphysics®: Finite element software for electrochemical analysis. a mini-review", *Electrochemistry Communications*, vol. 40, pp. 71–74, 2014.

[82] J. L. Loeppky, J. Sacks, and W. J. Welch, "Choosing the sample size of a computer experiment: A practical guide", *Technometrics*, vol. 51, no. 4, pp. 366–376, 2009.

[83] D. Sicard, A. J. Haak, K. M. Choi, A. R. Craig, L. E. Fredenburgh, and D. J. Tschumperlin, "Aging and anatomical variations in lung tissue stiffness", *American Journal of Physiology-Lung Cellular and Molecular Physiology*, vol. 314, no. 6, pp. L946–L955, 2018.

[84] K. Liao, C. R. Schultesiz, D. L. Hunston, and L. C. Brinson, "Long-term durability of fiber-reinforced polymer-matrix composite materials for infrastructure applications: A review", *Journal of Advanced Materials*, vol. 30, no. 4, pp. 3–40, 1998.

[85] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity", *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[86] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *nature*, vol. 521, no. 7553, p. 436, 2015.

[87] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives", *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[88] Y. LeCun, K. Kavukcuoglu, C. Farabet, *et al.*, "Convolutional networks and applications in vision.", in *ISCAS*, vol. 2010, 2010, pp. 253–256.

[89] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database", in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Ieee, 2009, pp. 248–255.

[90] J. Wang, Y. Ma, L. Zhang, R. X. Gao, and D. Wu, "Deep learning for smart manufacturing: Methods and applications", *Journal of Manufacturing Systems*, vol. 48, pp. 144–156, 2018.

[91] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis", *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[92] S. S. Du, Y. Wang, X. Zhai, S. Balakrishnan, R. Salakhutdinov, and A. Singh, "How many samples are needed to learn a convolutional neural network?", *stat*, vol. 1050, p. 21, 2018.

[93] B. F. Stewart, D. Siscovick, B. K. Lind, J. M. Gardin, J. S. Gottdiener, V. E. Smith, D. W. Kitzman, C. M. Otto, *et al.*, "Clinical factors associated with calcific aortic valve disease", *Journal of the American College of Cardiology*, vol. 29, no. 3, pp. 630–634, 1997.

[94] J. D. Anderson and J. Wendt, *Computational fluid dynamics*. Springer, 1995, vol. 206.

[95] P. A. Thompson and S. M. Troian, "A general boundary condition for liquid flow at solid surfaces", *Nature*, vol. 389, no. 6649, p. 360, 1997.

[96] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[97] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets", in *Proceedings of the British Machine Vision Conference*, BMVA Press, 2014.

[98] J. Wang and L. Perez, "The effectiveness of data augmentation in image classification using deep learning", *Convolutional Neural Networks Vis. Recognit*, p. 11, 2017.

[99] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks", in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[100] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[101] D. P. Kingma and M. Welling, "Auto-encoding variational bayes", *stat*, vol. 1050, p. 1, 2014.

[102] S. Zhao, J. Song, and S. Ermon, "Towards deeper understanding of variational autoencoding models", *arXiv preprint arXiv:1702.08658*, 2017.

[103] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions", in *Advances in Neural Information Processing Systems*, 2018, pp. 10 215–10 224.

[104] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville, "Adversarially learned inference", *stat*, vol. 1050, p. 2, 2016.

[105] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning", *arXiv preprint arXiv:1605.09782*, 2016.

[106] M. Mirza and S. Osindero, "Conditional generative adversarial nets", *arXiv preprint arXiv:1411.1784*, 2014.

[107] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans", in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, 2017, pp. 2642–2651.

[108] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification", in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, IEEE, 2018, pp. 289–293.

[109] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?", in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[110] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey", *IEEE transactions on neural networks and learning systems*, vol. 26, no. 5, pp. 1019–1034, 2014.

[111] B. J. Winer, "Statistical principles in experimental design.", 1962.

[112] B. Settles, "Active learning literature survey", University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[113] ——, "Active learning", *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012.

[114] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data", in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, 2017, pp. 1183–1192.

[115] M. Ducoffe and F. Precioso, "Adversarial active learning for deep networks: A margin based approach", *arXiv preprint arXiv:1802.09841*, 2018.

[116] J.-J. Zhu and J. Bento, "Generative adversarial active learning", *arXiv preprint arXiv:1702.07956*, 2017.

[117] Y. Wang and Q. Yao, "Few-shot learning: A survey", *arXiv preprint arXiv:1904.05046*, 2019.

[118] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks", in *International Conference on Machine Learning*, 2017, pp. 214–223.

[119] S. I. Resnick, *A Probability Path*. Springer Science & Business Media, 2013.

[120] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[121] C. E. Shannon, "A mathematical theory of communication", *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[122] V. R. Joseph, T. Dasgupta, R. Tuo, and C. J. Wu, "Sequential exploration of complex surfaces using minimum energy designs", *Technometrics*, vol. 57, no. 1, pp. 64–74, 2015.

[123] S. Mak, V. R. Joseph, *et al.*, "Support points", *The Annals of Statistics*, vol. 46, no. 6A, pp. 2562–2592, 2018.

[124] E. A. Nadaraya, "On estimating regression", *Theory of Probability & its Applications*, vol. 9, no. 1, pp. 141–142, 1964.

[125] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[126] H. Xiao, K. Rasul, and R. Vollgraf. (Aug. 28, 2017). "Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms". arXiv: cs.LG/1708.07747 [cs.LG].

[127] Y. LeCun, "The mnist database of handwritten digits", *http://yann. lecun. com/exdb/mnist/*, 1998.

[128] J. Chen, S. Mak, V. R. Joseph, and C. Zhang, "Adaptive design for Gaussian process regression under censoring", *arXiv preprint arXiv:1910.05452*, 2019.

[129] Zalando Research, *Fashion mnist*, https://github.com/zalandoresearch/fashion-mnist, 2017.

[130] S. Zagoruyko and N. Komodakis, "Wide residual networks", *arXiv preprint arXiv:1605.07146*, 2016.

[131] S. J. Pan, Q. Yang, *et al.*, "A survey on transfer learning", *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[132] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations", *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[133] S. U. Kim and J. De Vellis, "Stem cell-based cell therapy in neurological diseases: A review", *Journal of Neuroscience Research*, vol. 87, no. 10, pp. 2183–2200, 2009.

[134] P. Yin, "Keys to scale-up CAR T-cell therapy manufacturing", *BioProcess Online*, 2017.

[135] C. L. Bonifant, H. J. Jackson, R. J. Brentjens, and K. J. Curran, "Toxicity and management in car t-cell therapy", *Molecular Therapy-Oncolytics*, vol. 3, p. 16 011, 2016.

[136] C. H. June, R. S. O'Connor, O. U. Kawalekar, S. Ghassemi, and M. C. Milone, "CAR T cell immunotherapy for human cancer", *Science*, vol. 359, no. 6382, pp. 1361–1365, 2018.

[137] V. Prasad, "Tisagenlecleucel – the first approved CAR-T-cell therapy: Implications for payers and policy makers", *Nature Reviews Clinical Oncology*, vol. 15, no. 1, pp. 11–12, 2018.

[138] C. S. Hinrichs and N. P. Restifo, "Reassessing target antigens for adoptive T-cell therapy", *Nature Biotechnology*, vol. 31, no. 11, p. 999, 2013.

[139] R. P. Harrison, E. Zylberberg, S. Ellison, and B. L. Levine, "Chimeric antigen receptor–T cell therapy manufacturing: Modelling the effect of offshore production on aggregate cost of goods", *Cytotherapy*, vol. 21, no. 2, pp. 224–233, 2019.

[140] C. Slouka, D. J. Wurm, G. Brunauer, A. Welzl-Wachter, O. Spadiut, J. Fleig, and C. Herwig, "A novel application for low frequency electrochemical impedance spectroscopy as an online process monitoring tool for viable cell concentrations", *Sensors*, vol. 16, no. 11, p. 1900, 2016.

[141] T. J. Collins, "ImageJ for microscopy", *Biotechniques*, vol. 43, no. S1, S25–S30, 2007.

[142] Y. Pan, N. Hu, X. Wei, L. Gong, B. Zhang, H. Wan, and P. Wang, "3D cell-based biosensor for cell viability and drug assessment by 3D electric cell/matrigel-substrate impedance sensing", *Biosensors and Bioelectronics*, vol. 130, pp. 344–351, 2019.

[143] E. Gheorghiu and K. Asami, "Monitoring cell cycle by impedance spectroscopy: Experimental and theoretical aspects", *Bioelectrochemistry and Bioenergetics*, vol. 45, no. 2, pp. 139–143, 1998.

[144] S. Goh and R. Ram, "Impedance spectroscopy for in situ biomass measurements in microbioreactor", in *Proceedings of the 14th International Conference on Miniaturized Systems for Chemistry and Life Science, Groningen, The Netherlands*, 2010, pp. 3–7.

[145] R. Tuo and C. F. J. Wu, "Efficient calibration for imperfect computer models", *The Annals of Statistics*, vol. 43, no. 6, pp. 2331–2352, 2015.

[146] T. Miura and S. Uno, "Computer simulation for electrochemical impedance of a living cell adhered on the inter-digitated electrode sensors", *Japanese Journal of Applied Physics*, vol. 58, no. SB, SBBG15, 2019.

[147] H. P. Schwan, "Electrical properties of tissue and cell suspensions", in *Advances in Biological and Medical Physics*, vol. 5, Elsevier, 1957, pp. 147–209.

[148] V. R. Joseph and S. N. Melkote, "Statistical adjustments to engineering models", *Journal of Quality Technology*, vol. 41, no. 4, pp. 362–375, 2009.

[149] M. Plumlee, V. R. Joseph, and H. Yang, "Calibrating functional parameters in the ion channel models of cardiac cells", *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 500–509, 2016.

[150] D. A. Brown and S. Atamturktur, "Nonparametric functional calibration of computer models", *Statistica Sinica*, pp. 721–742, 2018.

[151] A. A. Ezzat, A. Pourhabib, and Y. Ding, "Sequential design for functional calibration of computer models", *Technometrics*, vol. 60, no. 3, pp. 286–296, 2018.

[152] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems*. Elsevier, 2018.

[153] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[154] J. W. Haycock, "3D cell culture: A review of current approaches and techniques", in *3D Cell Culture*, Springer, 2011, pp. 1–15.

[155] J. O. Ramsay, "Functional data analysis", *Encyclopedia of Statistical Sciences*, vol. 4, 2004.

[156] J.-L. Wang, J.-M. Chiou, and H.-G. Müller, "Functional data analysis", *Annual Review of Statistics and its Application*, vol. 3, pp. 257–295, 2016.

[157] G. E. P. Box and D. R. Cox, "An analysis of transformations", *Journal of the Royal Statistical Society, Series B*, vol. 26, pp. 211–252, 1964.

[158] R. M. Sakia, "The Box-Cox transformation technique: A review", *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 41, no. 2, pp. 169–178, 1992.

[159] I.-K. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry", *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000.

[160] R. Tibshirani, "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[161] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution", *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 6, pp. 797–829, 2006.

[162] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[163] P. Dlugolecki, P. Ogonowski, S. J. Metz, M. Saakes, K. Nijmeijer, and M. Wessling, "On the resistances of membrane, diffusion boundary layer and double layer in ion exchange membrane transport", *Journal of Membrane Science*, vol. 349, no. 1-2, pp. 369–379, 2010.

[164] L.-H. Lin and V. R. Joseph, "Transformation and additivity in gaussian processes", *Technometrics*, pp. 1–11, 2019.

[165] J. Chen, Z. Chen, C. Zhang, and C. F. J. Wu, "APIK: Active physics-informed kriging model with partial differential equations", *arXiv preprint arXiv:2012.11798*, 2020.

[166] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations", *arXiv preprint arXiv:1711.10561*, 2017.